JOHANNES GUTENBERG UNIVERSITÄT MAINZ

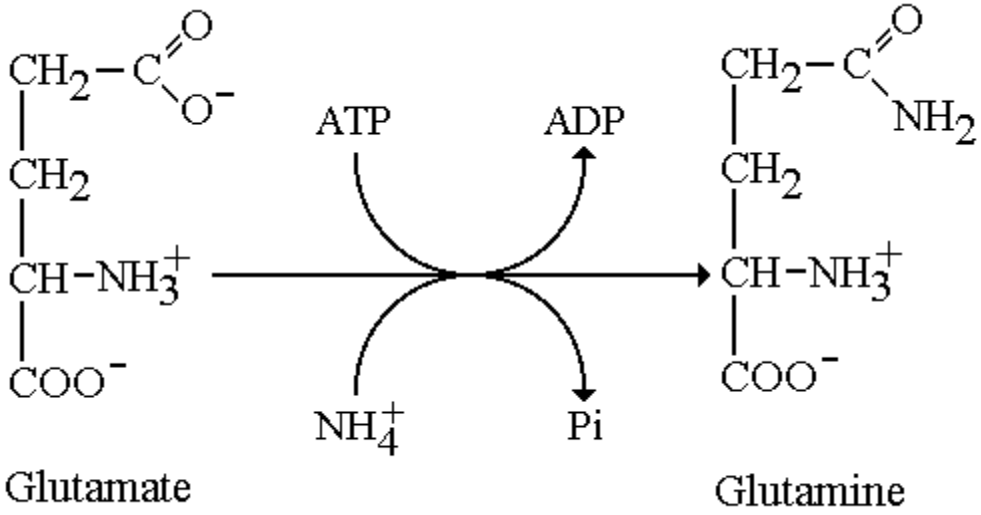
Homology 3D modeling

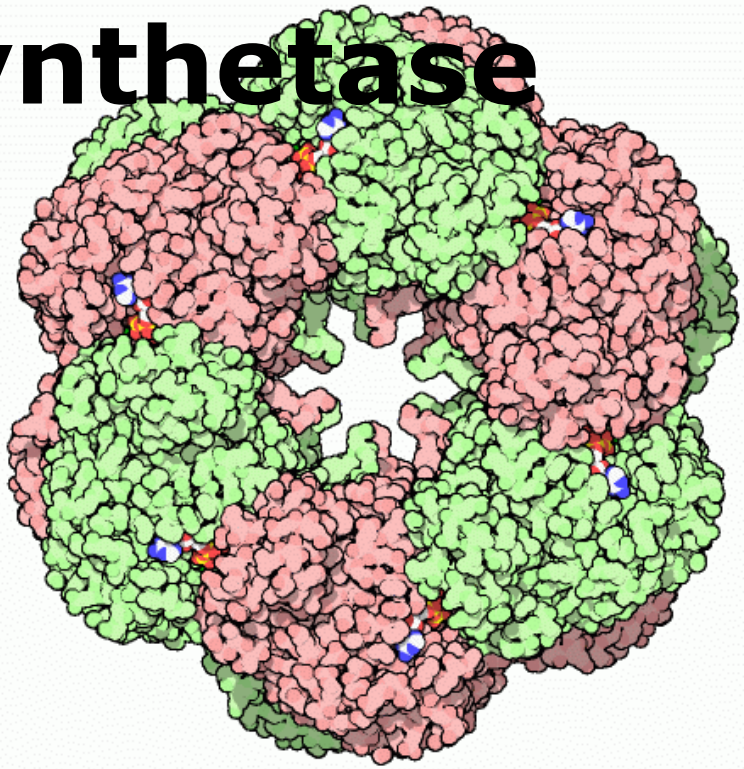
Miguel Andrade
Faculty of Biology,
Institute of Organismic and Molecular Evolution
Johannes Gutenberg University
Mainz, Germany
andrade@uni-mainz.de

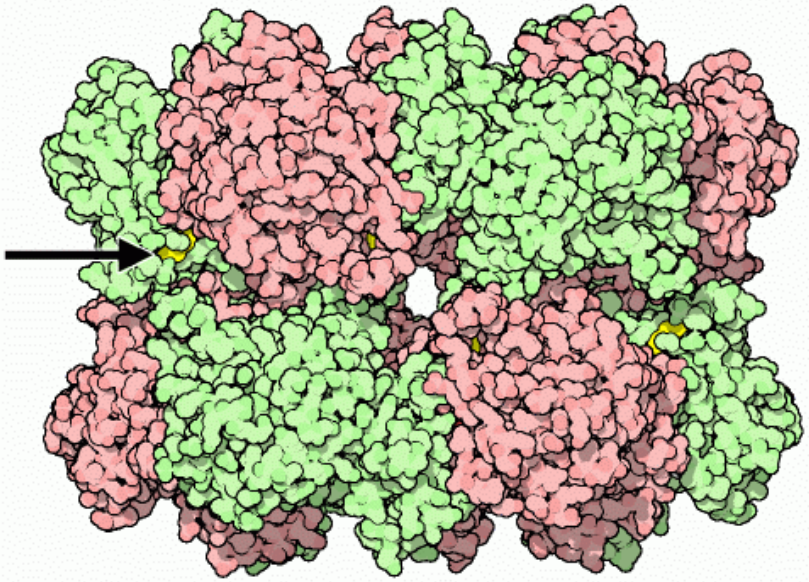
Mount Everest

Age: 60M years

Glutamine synthetase



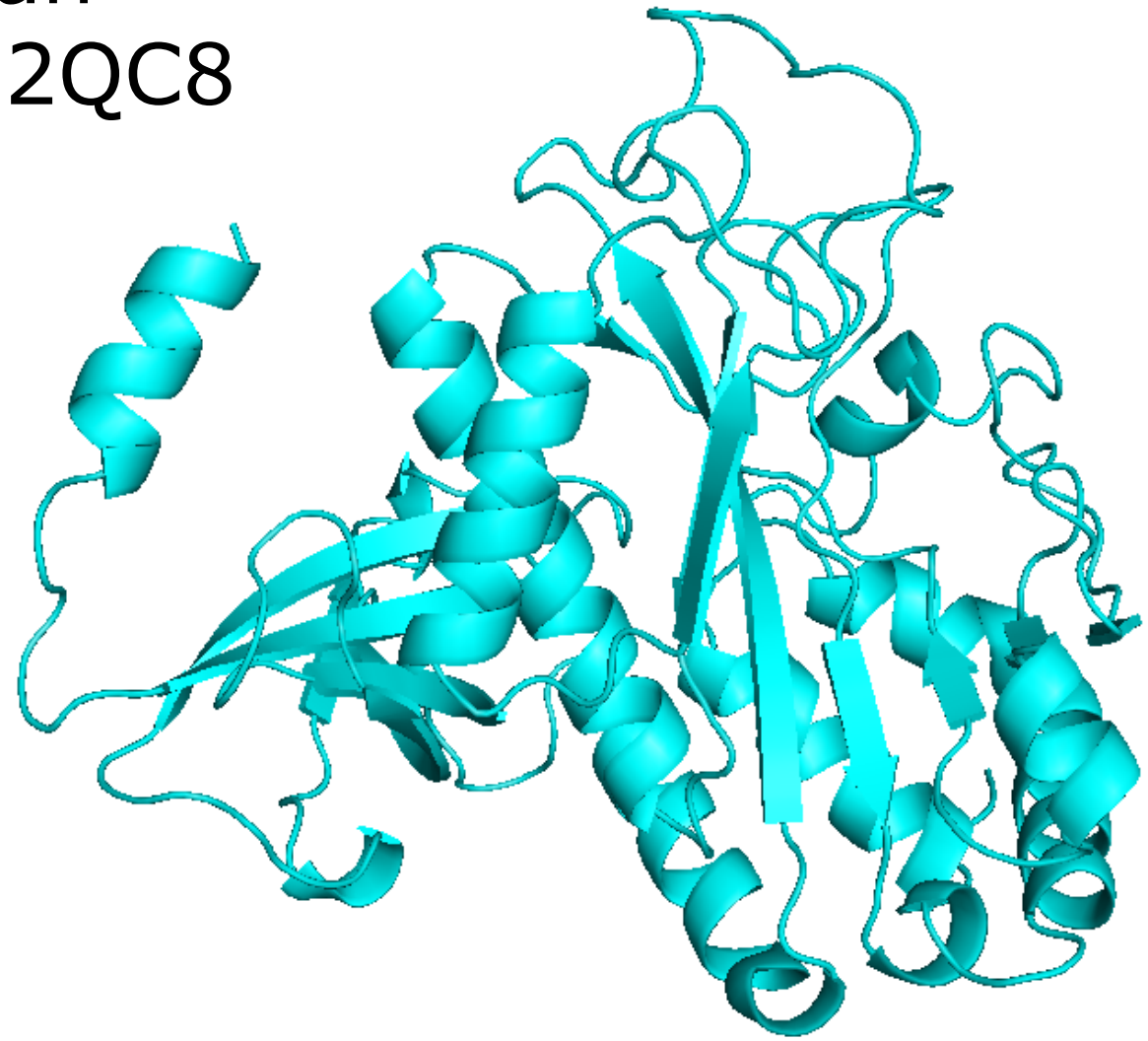




Age: +3500M years

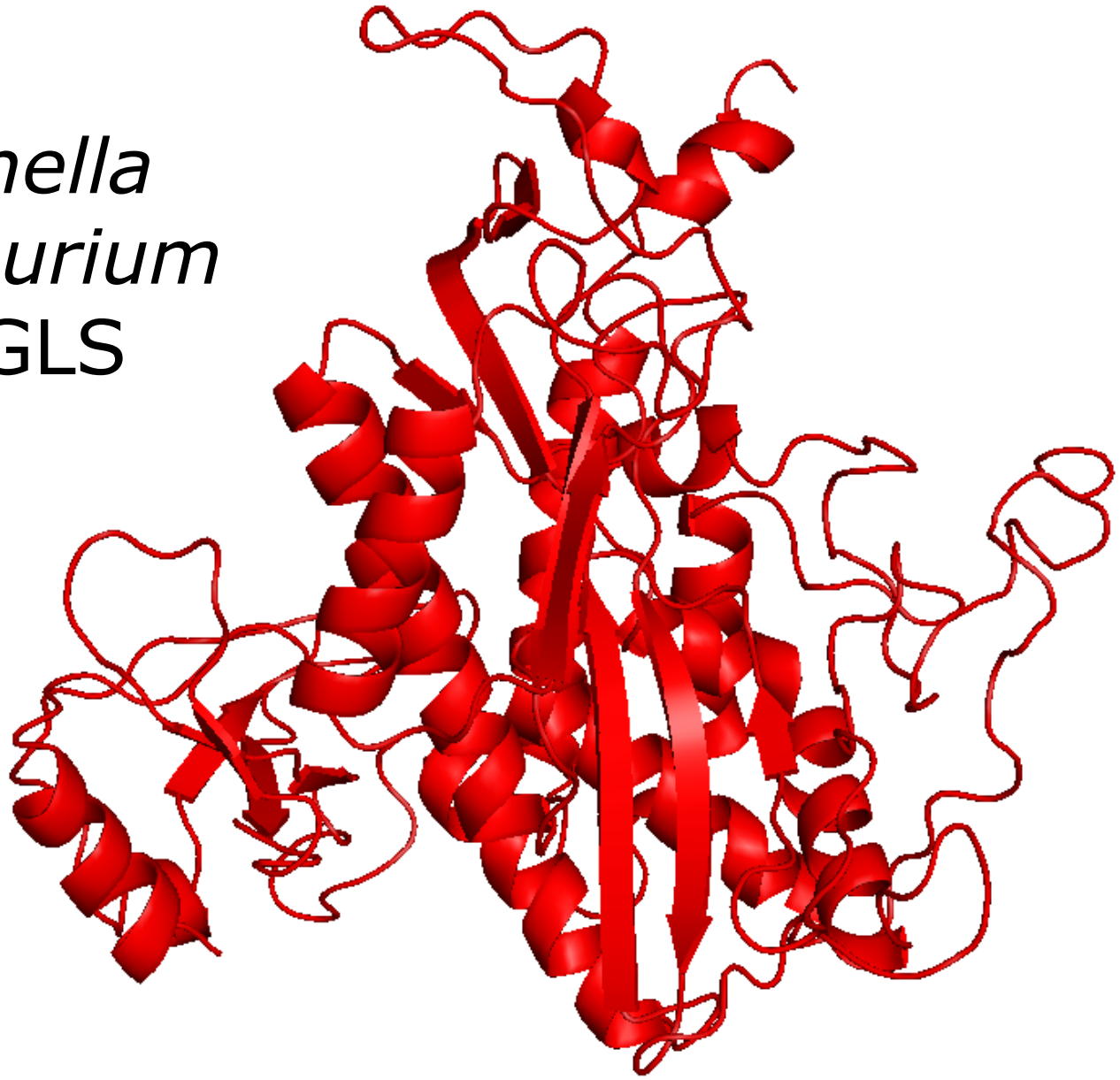
Glutamine synthetase

Human
PDB:2QC8

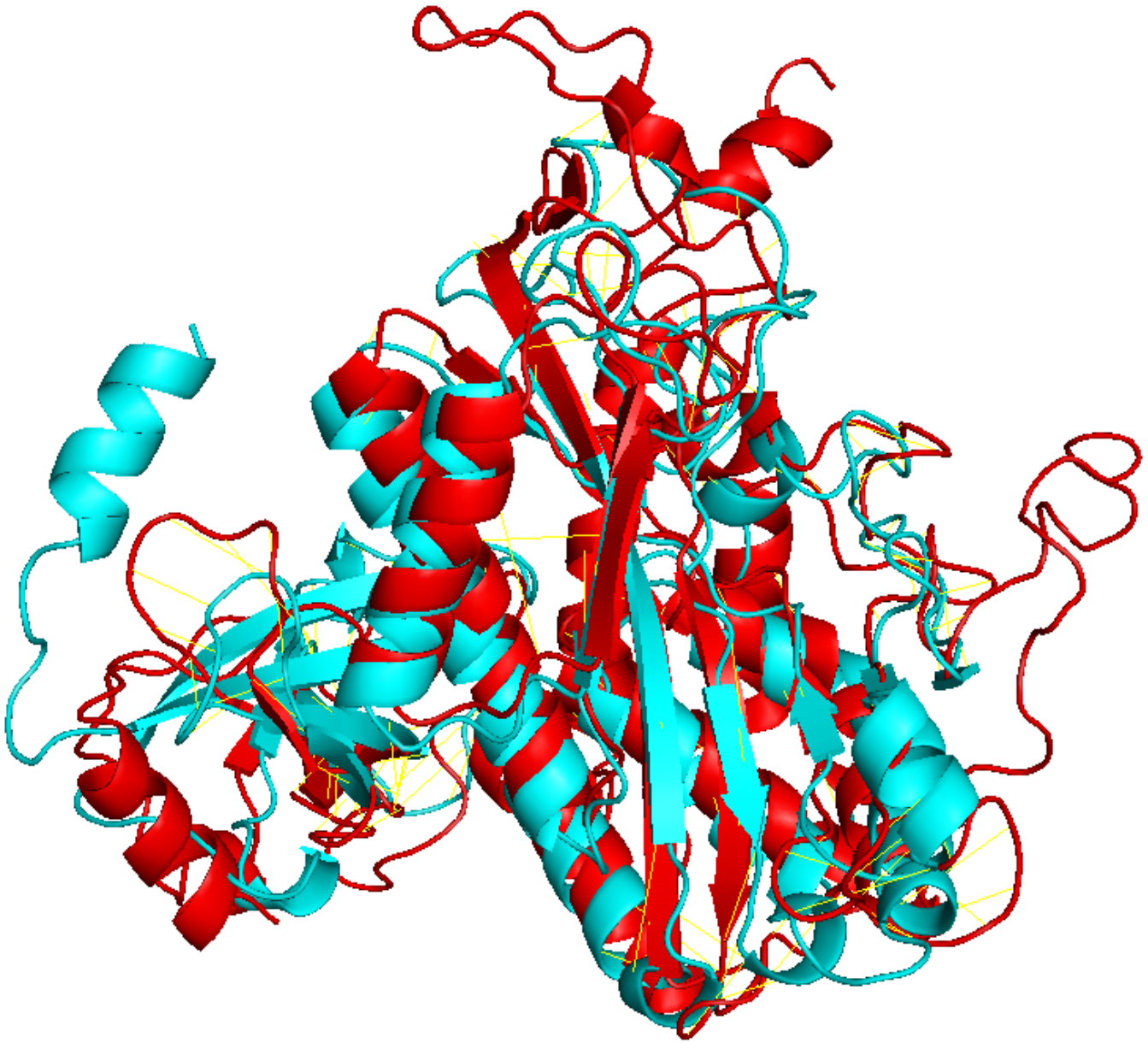


Glutamine synthetase

Salmonella typhimurium
PDB:2GLS



Glutamine synthetase



Time line

Earth: 4.6 By

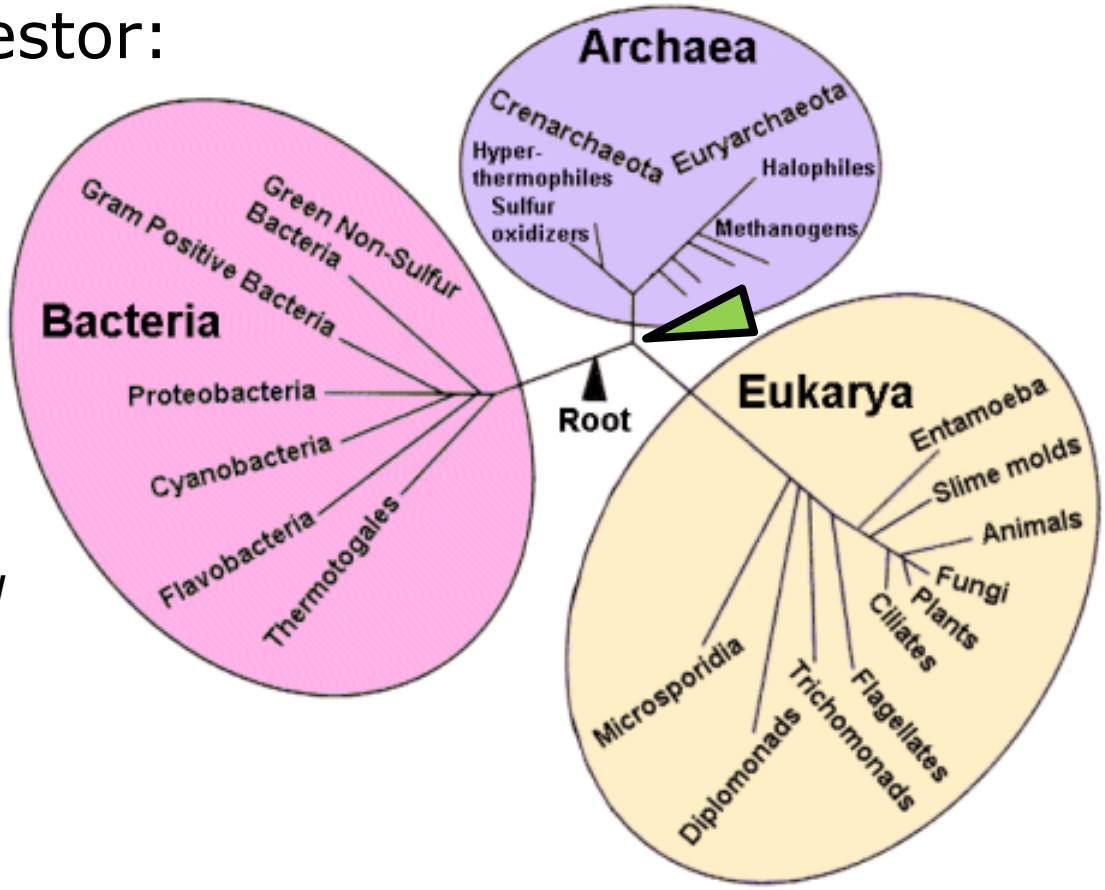
Origin of life: 3.9 By – 3.5 By

Last Common Ancestor:
3.5 – 3.8 By

Glansdorff & Labedan (2008) *Biology Direct*

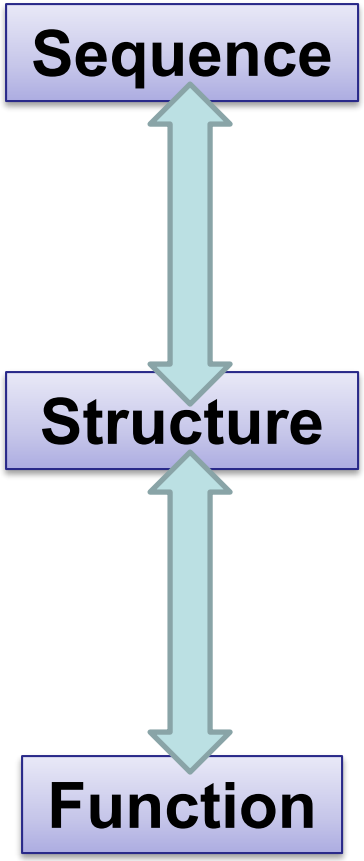
4.29 By

Sheridan *et al.* (2003) *Geomicrobiology Journal*



Sequence and function

Evolutionary constraints



MTQDELKKAVGWAALQYVQ

PG DA

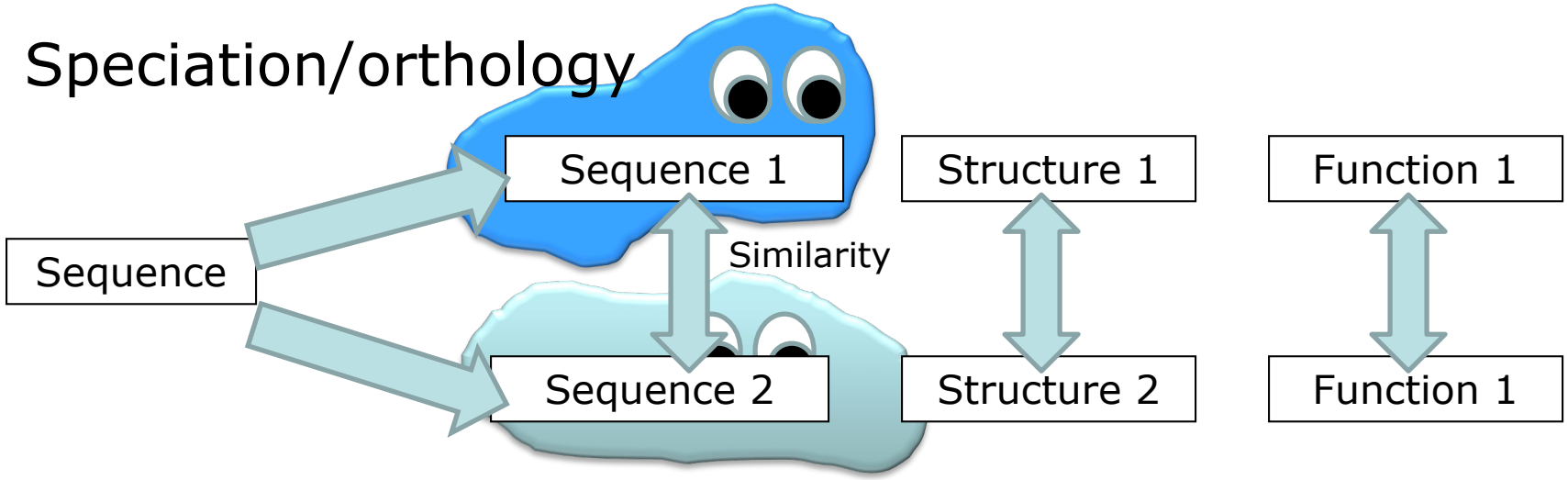
LG ST

EK

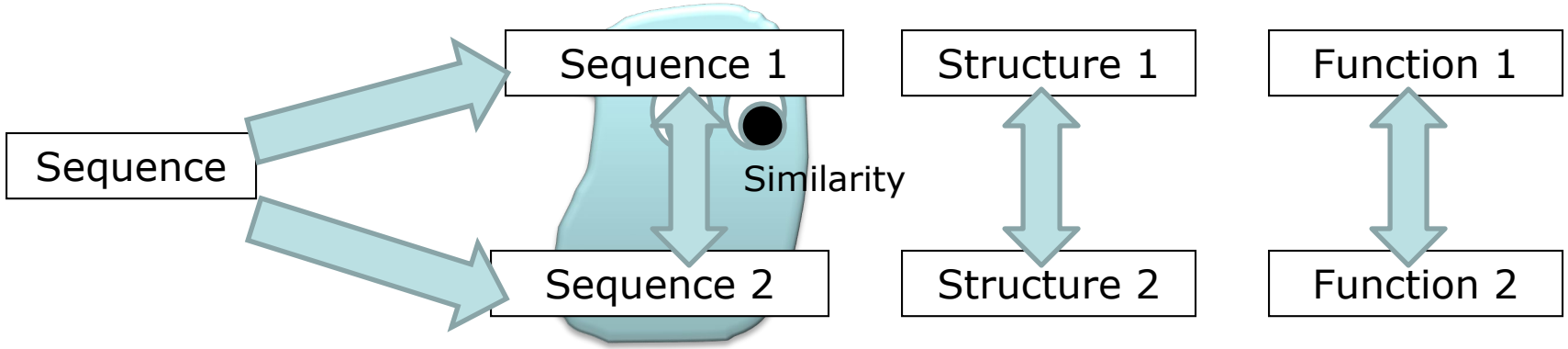
Sequence and function

Evolutionary constraints

Speciation/orthology



Gene duplication/paralogy



Sequence pairwise alignment

>gs_human gi|74271837|ref|NP_001028216.1| glutamine synthetase [Homo sapiens]
MTTSASSHLNKGIKQVYMSLPQGEKVQAMYIWIDGTGEGLRCKTRTLDSEPKCVEELPEWNFDGSSTLQS
EGSNSDMYLVPAAMFRDPFRKDPNKLVLCEVFKYNRRPAETNLRHTCKRIMDMVSNQHPWFGMEQEYTLM
GTDGHPFGWPSNGFPGPQGPYYCGVGADRAYGRDIVEAHYRACLYAGVKIAGTNAEVMPAQWEFQIGPCE
GISMGDHLWVARFILHRVCEDFGVIATFDPKPIPGNWNGAGCHTNFSTKAMREENGLKYIEEAIEKLSKR
HQYHIRAYDPKGGLDNARRLTGFHETSNINDFSAGVANRSASIRIPRTVGQEKKGYFEDRRPSANCDPFS
VTEALIRTCLLNETGDEPFQYKN

>gs_salmonella gi|16767272|ref|NP_462887.1| glutamine synthetase [Salmonella enterica subsp. enterica serovar Typhimurium str. LT2]
MSAEHVLTMLNEHEVKFVDLRFTDTKGKEQHVTIPAHQVNAEFFEEGKMFDGSSIGGWKGINESDMVLMP
DASTAVIDPFFADSTLIIRCDILEPGTLQGYDRDPRSIAKRAEDYLRATGIADTVLFGPEPEFFLFDDIR
FGASISGSHVAIDDIEGAWNSSTKYEGGNKGHRPGVKGGYFPVPPVDSAQDIRSEMCLVMEQMGLVVEAH
HHEVATAGQNEVATRFNTMTKKADEIQIYKYVVHNVAHRFGKTATFMPKPMFGDNGSGMHCHMSLAKNGT
NLFSGDKYAGLSEQALYYIGGVIKHAKAINALANPTTNSYKRLVPGYEAPVMLAYSARNRSASIRIPVVA
SPKARRIEVRFPDPAANPYLCFAALLMAGLDGIKNKIHPGEAMDKNLYDLPPEEAKEIPQVAGSLEEALN
ALDLDREFLKAGGVFTDEAIDAYIALRREEDDRVRMTPHPVEFELYYSV

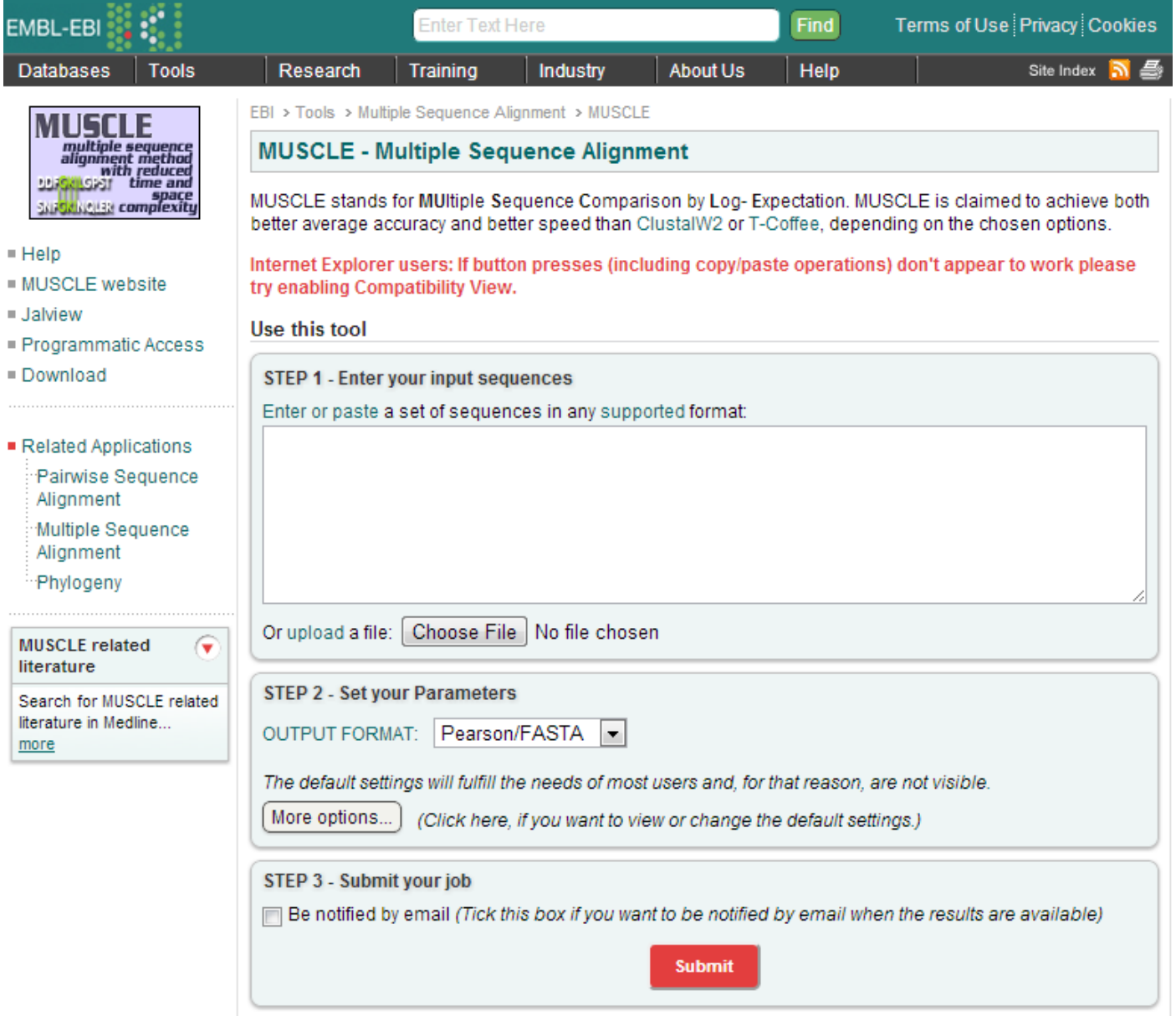
Sequence pairwise alignment

BLAST (Altschul et al, 1990)

```
>lcl|39919 unnamed protein product
Length=469

 Score = 70.5 bits (171),  Expect = 1e-17, Method: Compositional matrix adjust.
 Identities = 102/363 (28%), Positives = 138/363 (38%), Gaps = 96/363 (26%)

Query  62   FDGSSTLQSEGSN-SDMYLVPAA--MFRDPFRKDPNKLVLCEVFK------YNRRP----  108
            FDGSS    +G N SDM L+P A     DPF  D   ++ C++ +      Y+R P
Sbjct  50   FDGSSIGGWKGINESDMVLMPDASTAVIDPFFADSTLIIRCDILEPGTLQGYDRDPRSIA  109

Query  109  --AETNLRHTCKRIMDMVSNQHPWFGMEQEYTLMGTDGHPFGWPSNGF-----------  154
              AE  LR T   I D V      FG E E+ L   D   FG   +G
Sbjct  110  KRAEDYLRATG--IADTV-----LFGPEPEFFLF--DDIRFGASISGSHVAIDDIEGAWN  160

Query  155  ------------PGPQGPYYCGVGADRAYGRDI-------------VEAHYRACLYAG  187
                        PG +G Y+     D A  +DI               VEAH+     AG
Sbjct  161  SSTKYEGGNKGHRPGVKGGYFPVPPVDSA--QDIRSEMCLVMEQMGLVVEAHHHEVATAG  218

Query  188  VKIAGTNAEVMPAQWEFQIGPCEGISMGDHLWVARFILHRVCEDFGVIATFDPKPIPG-N  246
                 T    M  +              D + + ++++H V   FG  ATF PKP+ G N
Sbjct  219  QNEVATRFNTMTKK------------ADEIQIYKYVVHNVAHRFGKTATFMPKPMFGDN  265

Query  247  WNGAGCHTNFSTKAMREENGLKYIEEAIEKLSKRHQYHIRAYDPKGGLDNA---------  297
              +G  CH + +       +G KY       LS++  Y+I           NA
Sbjct  266  GSGMHCHMSLAKNGTNLFSGDKY-----AGLSEQALYYIGGVIKHAKAINALANPTTNSY  320

Query  298  RRLTGFHETSNINDFSAGVANRSASIRIPRTVGQEKKGYFEDRRPSANCDPFSVTEALIR  357
            +RL   +E   +  +SA   NRSASIRIP  V   K    E R P    +P+    AL+
Sbjct  321  KRLVPGYEAPVMLAYSA--RNRSASIRIP-VVASPKARRIEVRFPDPAANPYLCFAALLM  377

Query  358  TCL  360
              L
Sbjct  378  AGL  380
```

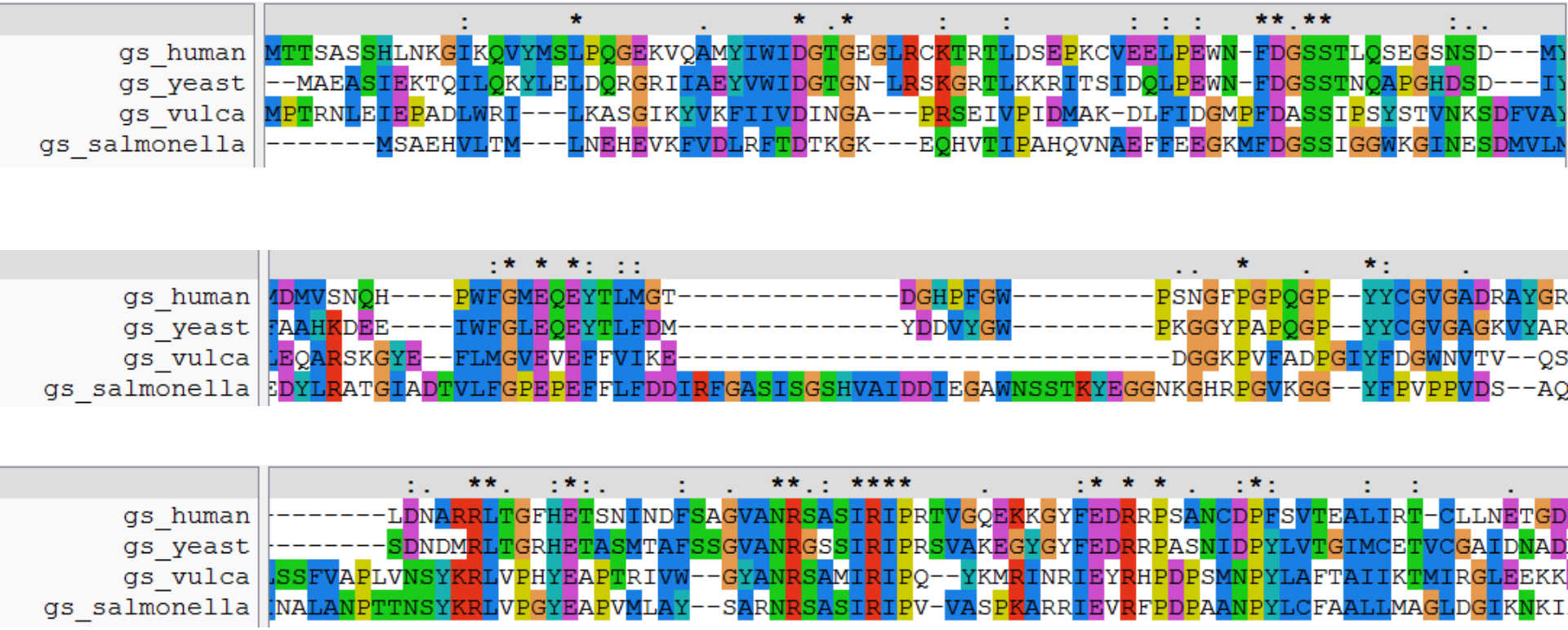
Multiple sequence alignment

>gs_human gi|74271837|ref|NP_001028216.1| glutamine synthetase [Homo sapiens]
MTTSASSHLNKGIKQVYMSLPQGEKVQAMYIWIDGTGEGLRCKTRTLDSEPKCVEELPEWNFDGSSTLQS
EGSNSDMYLVPAAMFRDPFRKDPNKLVLCEVFKYNRRPAETNLRHTCKRIMDMVSNQHPWFGMEQEYTLM
GTDGHPFGWPSNGFPGPQGPYYCGVGADRAYGRDIVEAHYRACLYAGVKIAGTNAEVMPAQWEFQIGPCE
GISMGDHLWVARFILHRVCEDFGVIATFDPKPIPGNWNGAGCHTNFSTKAMREENGLKYIEEAIEKLSKR
HQYHIRAYDPKGGLDNARRLTGFHETSNINDFSAGVANRSASIRIPRTVGQEKKGYFEDRRPSANCDPFS
VTEALIRTCLLNETGDEPFQYKN
>gs_vulca gi|307594850|ref|YP_003901167.1| glutamine synthetase [Vulcanisaeta
distributa DSM 14429]
MPTRNLEIEPADLWRILKASGIKYVKFIIVDINGAPRSEIVPIDMAKDLFIDGMPFDASSIPSYSTVNKS
DFVAYVDPRAVYVEYWQDGKVADVFTMVSDIADKPSPLDPRRVLNDALEQARSKGYEFLMGVEVEFFVIK
EDGGKPVFADPGIYFDGWNVTVQSQFMKELITAIADAGINYTKTHHEVAPSQYEVNIGATDPLRLADQIV
YFKIMAKDIARKYGLVATFMPKPFWGVNGSGAHTHISVWKDGKNLFQSSTGKITEECGYAISAILSNARA
LSSFVAPLVNSYKRLVPHYEAPTRIVWGYANRSAMIRIPQYKMRINRIEYRHPDPSMNPYLAFTAIIKTM
IRGLEEKKEPPPPTEEVAYELANALETPATLEDTLKELSKSFLATELPSELVNAYIKIKQNEWEDYLTNV
GPWEKTWNIITQWEYNKYLVTA
>gs_salmonella gi|16767272|ref|NP_462887.1| glutamine synthetase [Salmonella
enterica subsp. enterica serovar Typhimurium str. LT2]
MSAEHVLTMLNEHEVKFVDLRFTDTKGKEQHVTIPAHQVNAEFFEEGKMFDGSSIGGWKGINESDMVLMP
DASTAVIDPFFADSTLIIRCDILEPGTLQGYDRDPRSIAKRAEDYLRATGIADTVLFGPEPEFFLFDDIR
FGASISGSHVAIDDIEGAWNSSTKYEGGNKGHRPGVKGGYFPVPPVDSAQDIRSEMCLVMEQMGLVVEAH
HHEVATAGQNEVATRFNTMTKKADEIQIYKYVVHNVAHRFGKTATFMPKPMFGDNGSGMHCHMSLAKNGT
NLFSGDKYAGLSEQALYYIGGVIKHAKAINALANPTTNSYKRLVPGYEAPVMLAYSARNRSASIRIPVVA
SPKARRIEVRFPDPAANPYLCFAALLMAGLDGIKNKIHPGEAMDKNLYDLPPEEAKEIPQVAGSLEEALN
ALDLDREFLKAGGVFTDEAIDAYIALRREEDDRVRMTPHPVEFELYYSV
>gs_yeast gi|330443748|ref|NP_015360.2| Gln1p [Saccharomyces cerevisiae S288c]
MAEASIEKTQILQKYLELDQRGRIIAEYVWIDGTGNLRSKGRTLKKRITSIDQLPEWNFDGSSTNQAPGH
DSDIYLKPVAYYPDPFRRGDNIVVLAACYNNDGTPNKFNHRHEAAKLFAAHKDEEIWFGLEQEYTLFDMY
DDVYGWPKGGYPAPQGPYYCGVGAGKVYARDMIEAHYRACLYAGLEISGINAEVMPSQWEFQVGPCTGID
MGDQLWMARYFLHRVAEEFGIKISFHPKPLKGDWNGAGCHTNVSTKEMRQPGGMKYIEQAIEKLSKRHAE
HIKLYGSDNDMRLTGRHETASMTAFSSGVANRGSSIRIPRSVAKEGYGYFEDRRPASNIDPYLVTGIMCE
TVCGAIDNADMTKEFERESS



http://www.ebi.ac.uk/Tools/msa/muscle/

```
>gs_human gi|74271837|ref|NP_001028216.1| glutamine synthetase [Homo sapiens]
MTTSASSHLNKGIKQVYMSLPQGEKVQAMYIWIDGTGEGLRCKTRTLDSEPKCVEELPEW
N-FDGSSTLQSEGSNSD---MYLVPAAMFRDPFRKDPNKLVLCEVFKYNRRPA-ETNLRH
TCKRIMDMVSNQH----PWFGMEQEYTLMGT--------------DGHPFGW--------
-PSNGFPGPQGP--YYCGVGADRAYGRDIVEAHYRACLYAGVKIAGTNAEVMPA-QWEFQ
IGPCEGISMGDHLWVARFILHRVCEDFGVIATFDPKPIPGNWNGAGCHTNFSTKAMREEN
GLKYIEEAIEKLSKRHQYHIRAYDPKGG---------LDNARRLTGFHETSNINDFSAGV
ANRSASIRIPRTVGQEKKGYFEDRRPSANCDPFSVTEALIRT-CLLNETGDEP-------
------------------------------------------------------------
--------------FQYKN-----

>gs_yeast gi|330443748|ref|NP_015360.2| Gln1p [Saccharomyces cerevisiae S288c]
--MAEASIEKTQILQKYLELDQRGRIIAEYVWIDGTGN-LRSKGRTLKKRITSIDQLPEW
N-FDGSSTNQAPGHDSD---IYLKPVAYYPDPFRRGDNIVVLAACYNNDGTPN-KFNHRH
EAAKLFAAHKDEE----IWFGLEQEYTLFDM--------------YDDVYGW--------
-PKGGYPAPQGP--YYCGVGAGKVYARDMIEAHYRACLYAGLEISGINAEVMPS-QWEFQ
VGPCTGIDMGDQLWMARYFLHRVAEEFGIKISFHPKPLKGDWNGAGCHTNVSTKEMRQPG
GMKYIEQAIEKLSKRHAEHIKLYG-------------SDNDMRLTGRHETASMTAFSSGV
ANRGSSIRIPRSVAKEGYGYFEDRRPASNIDPYLVTGIMCETVCGAIDNADMT-------
------------------------------------------------------------
--------------KEFERESS—

>gs_vulca gi|307594850|ref|YP_003901167.1| glutamine synthetase [Vulcanisaeta distributa DSM 14429]
MPTRNLEIEPADLWRI---LKASGIKYVKFIIVDINGA---PRSEIVPIDMAK-DLFIDG
MPFDASSIPSYSTVNKSDFVAYVDPRAVYVEYWQDGKVADVFTMVSDIADKPS-PLDPRR
VLNDALEQARSKGYE--FLMGVEVEFFVIKE----------------------------
--DGGKPVFADPGIYFDGWNVTV--QSQFMKELITAIADAGINYTKTHHEVAPS-QYEVN
IGATDPLRLADQIVYFKIMAKDIARKYGLVATFMPKPFWGV-NGSGAHTHIS---VWKDG
KNLF-QSSTGKITEECGYAISAILSNARALSSFVAPLVNSYKRLVPHYEAPTRIVW--GY
ANRSAMIRIPQ--YKMRINRIEYRHPDPSMNPYLAFTAIIKTMIRGLEEKKEPPPPTEEV
AYELA--NALETP---ATLEDTLK--ELSKSFLATE--LPSELVNAYIKIKQNEWEDYLT
NVGPWEKTWNIITQWEYNKYLVTA

>gs_salmonella gi|16767272|ref|NP_462887.1| glutamine synthetase [Salmonella enterica
-------MSAEHVLTM---LNEHEVKFVDLRFTDTKGK---EQHVTIPAHQVNAEFFEEG
KMFDGSSIGGWKGINESDMVLMPDASTAVIDPFFADSTLIIRCDILEPGTLQGYDRDPRS
IAKRAEDYLRATGIADTVLFGPEPEFFLFDDIRFGASISGSHVAIDDIEGAWNSSTKYEG
GNKGHRPGVKGG--YFPVPPVDS--AQDIRSEMCLVMEQMGLVVEAHHHEVATAGQNEVA
TRFNTMTKKADEIQIYKYVVHNVAHRFGKTATFMPKPMFGD-NGSGMHCHMS---LAKNG
TNLFSGDKYAGLSEQALYYIGGVIKHAKAINALANPTTNSYKRLVPGYEAPVMLAY--SA
RNRSASIRIPV-VASPKARRIEVRFPDPAANPYLCFAALLMAGLDGIKNKIHPGEAMDKN
LYDLPPEEAKEIPQVAGSLEEALNALDLDREFLKAGGVFTDEAIDAYIALRREEDDRVRM
TPHP----------VEFELYYSV-
```



ClustalW, JalView

Determination of protein structure

X-ray crystallography (111K in PDB)

- need crystals

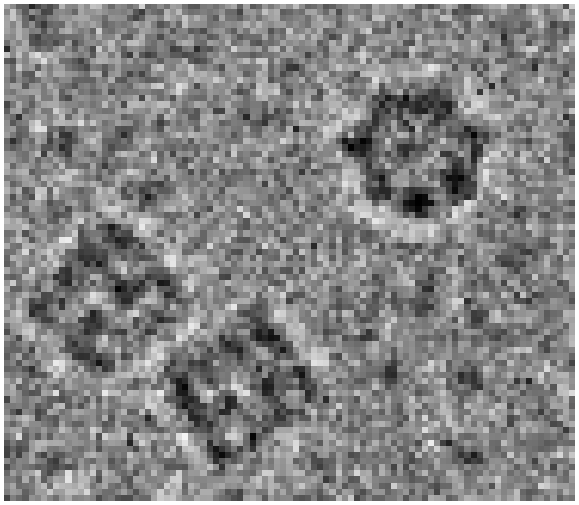
Nuclear Magnetic Resonance (NMR) (10.4K)

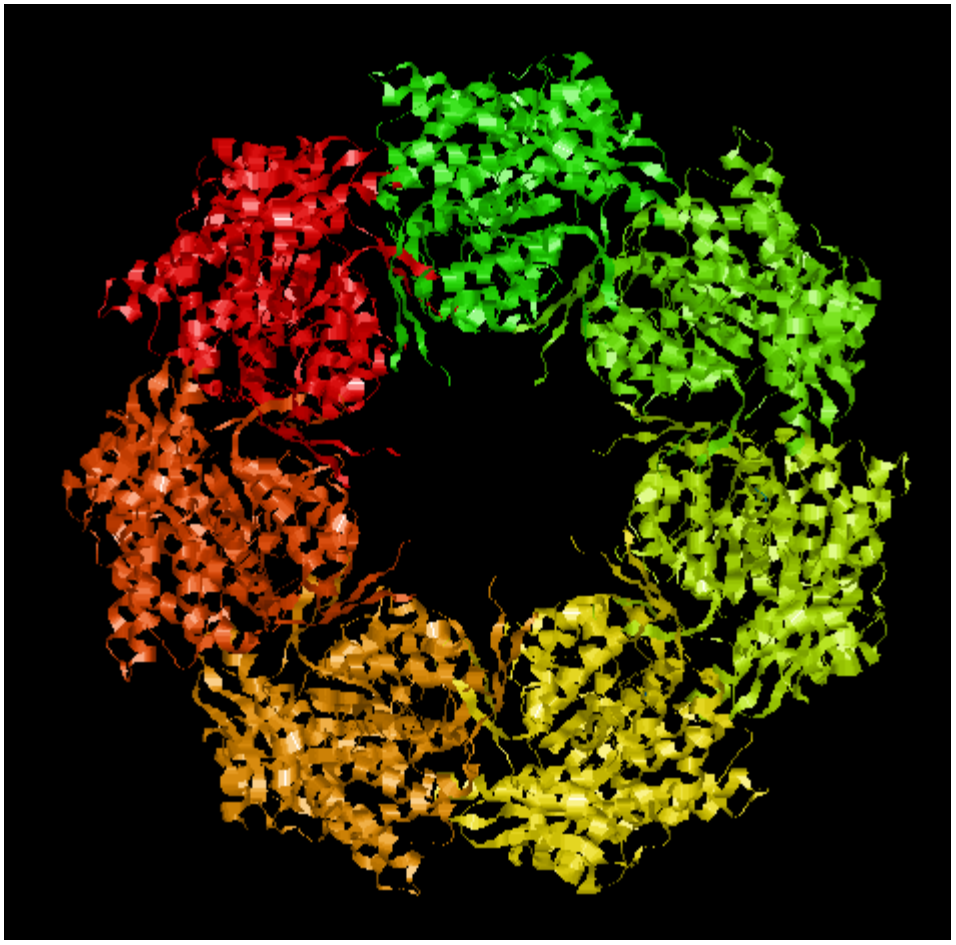
- proteins in solution
- lower size limit (600 aa)

Electron microscopy (1.2K)

- Low resolution (>5A)

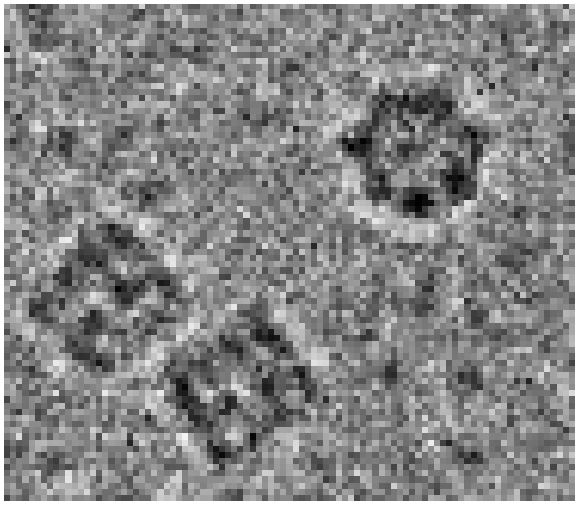
Determination of protein structure

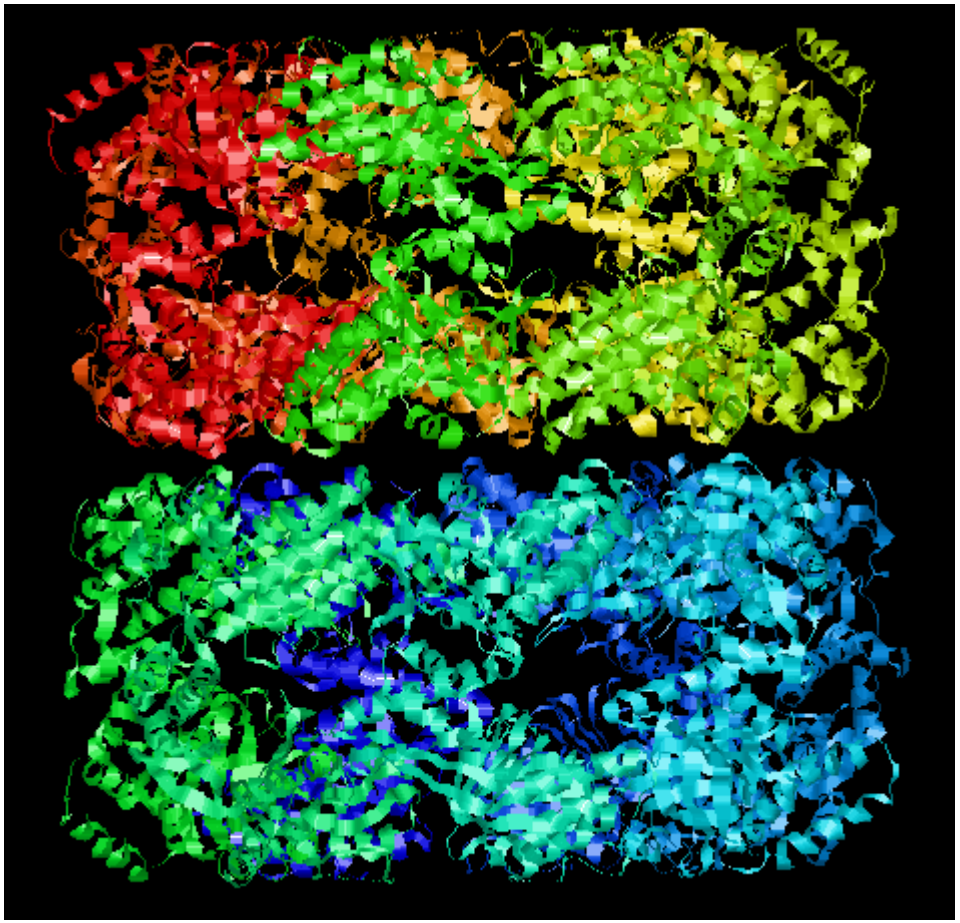




resolution 2.4 A

Determination of protein structure

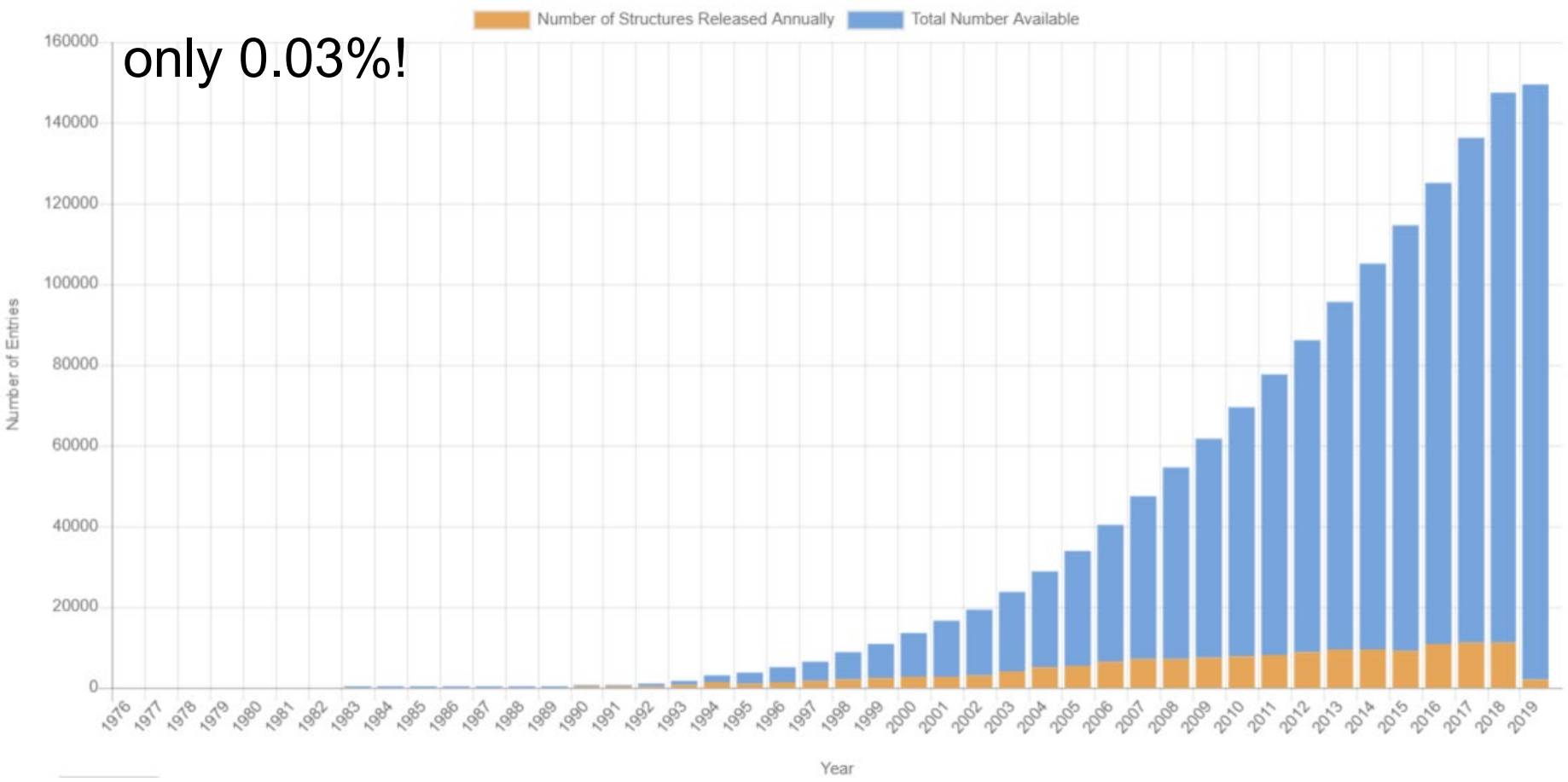




resolution 2.4 A

Structural genomics

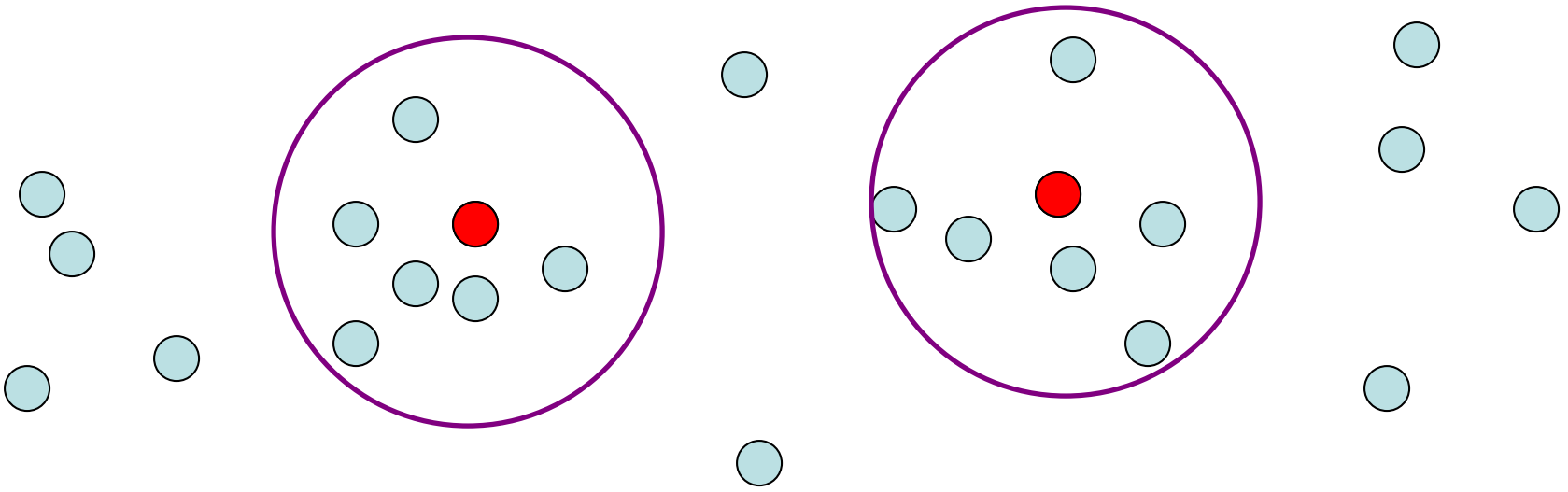
Currently: 150K protein 3D structures
from around 46.4K sequences in UniProt (how do I know?)
146M sequences in UniProt



Structural genomics

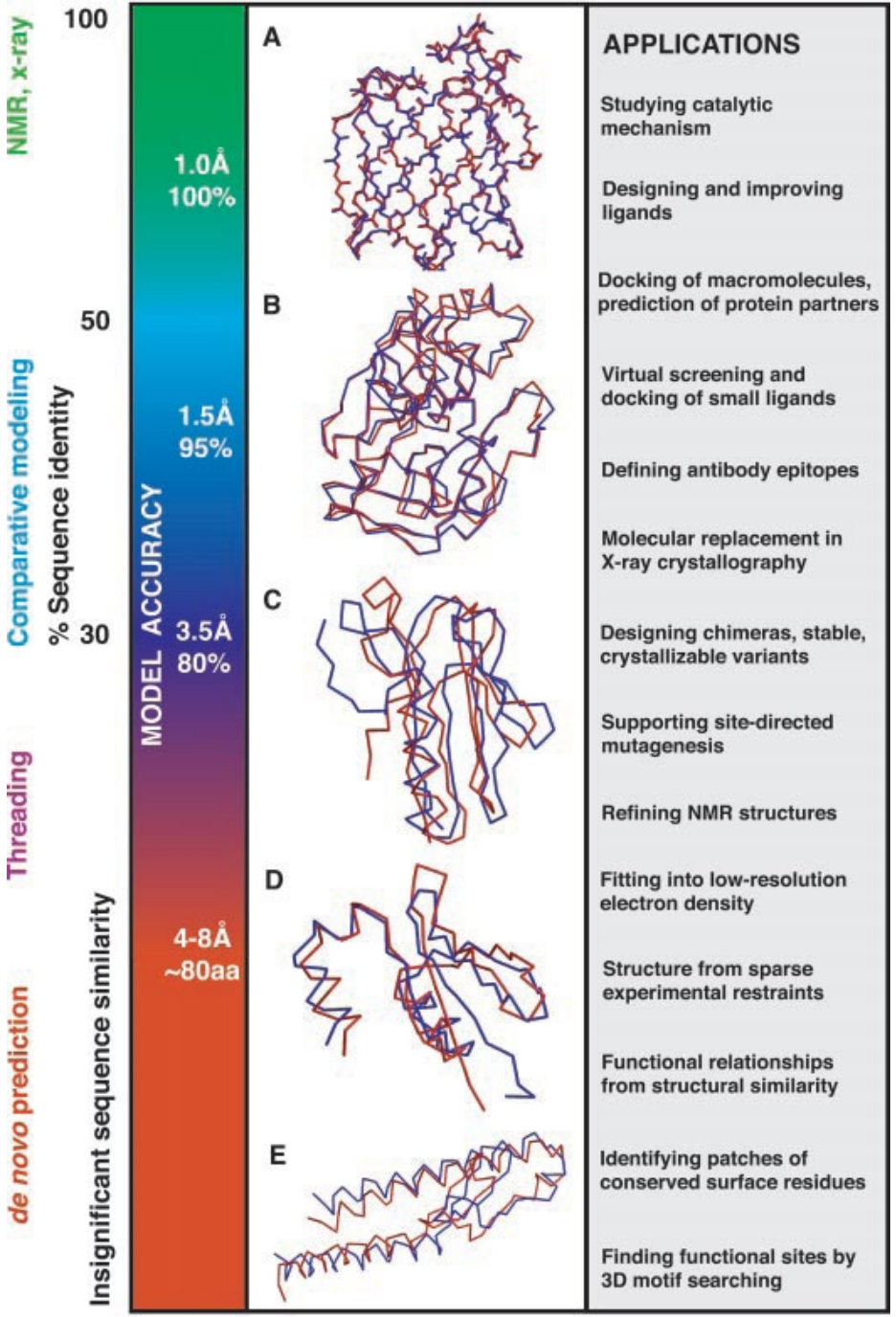
Currently: 150K protein 3D structures
from around 46.4K sequences in UniProt (how do I know?)
146M sequences in UniProt

only 0.03%!



50% sequences covered (25% in 1995)

Relation between sequence identity and accuracy/applications

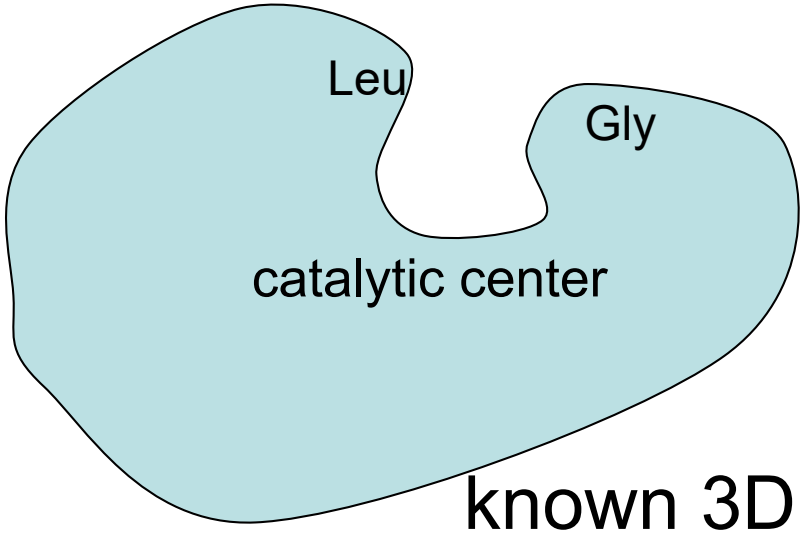


From:
Baker and Sali (2001)
Science

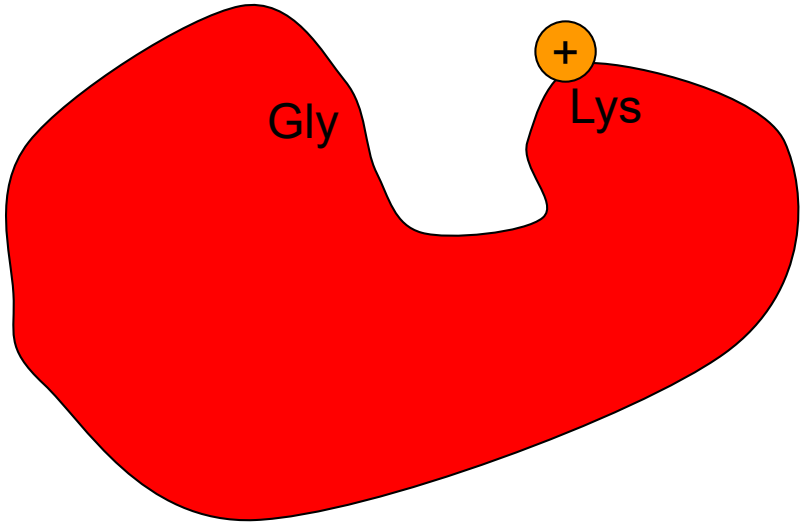
Homology modelling
Applications: target design

Query sequence

similar to



known 3D

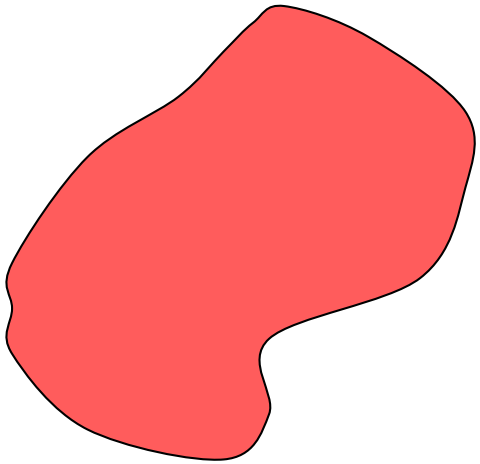


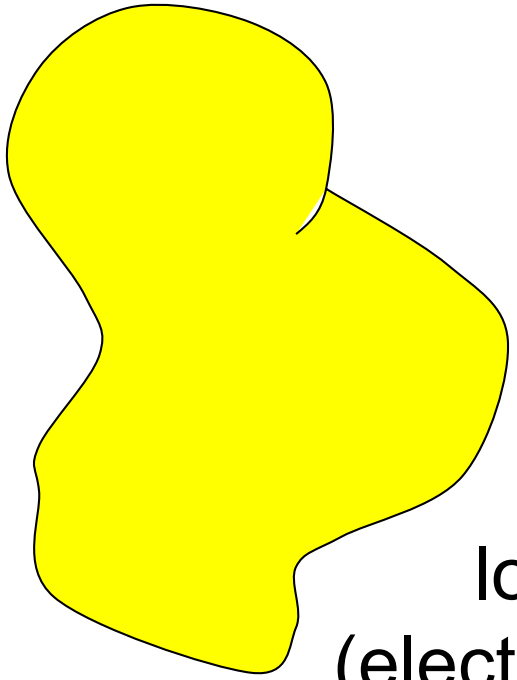
model 3D by homology

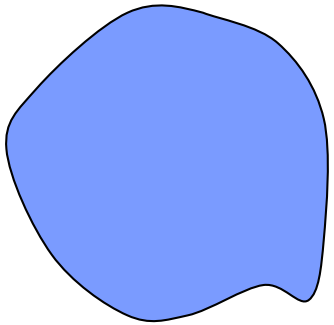
Homology modelling
Applications: fit to low res 3D

Query sequence 1

Query sequence 2







low resolution 3D
(electron microscopy)

Homology modelling

GenTHREADER

David Jones http://bioinf.cs.ucl.ac.uk/psipred_new/

Input sequence or MSA

Choose Prediction Methods

- [ ] PSIPRED v3.3 (Predict Secondary Structure)
- [x] pGenTHREADER (Profile Based Fold Recognition)
- [ ] BioSerf v2.0 (Automated Homology Modelling)
- [ ] FFPred v2.0 (Eukaryotic Function Prediction)
- [ ] MEMPACK (SVM Prediction of TM Topology and Helix Packing)
- [ ] DomSerf v2.0 (Automated Domain Modelling by Homology)
- [ ] DISOPRED3 & DISOPRED2 (Disorder Prediction)
- [ ] MEMSAT3 & MEMSAT-SVM (Membrane Helix Prediction)
- [ ] DomPred (Protein Domain Prediction)
- [ ] GenTHREADER (Rapid Fold Recognition)
- [ ] pDomTHREADER (Fold Domain Recognition)

Help...

Input Sequence (Single sequence or Multiple Sequence alignments; as raw sequence or fasta format)

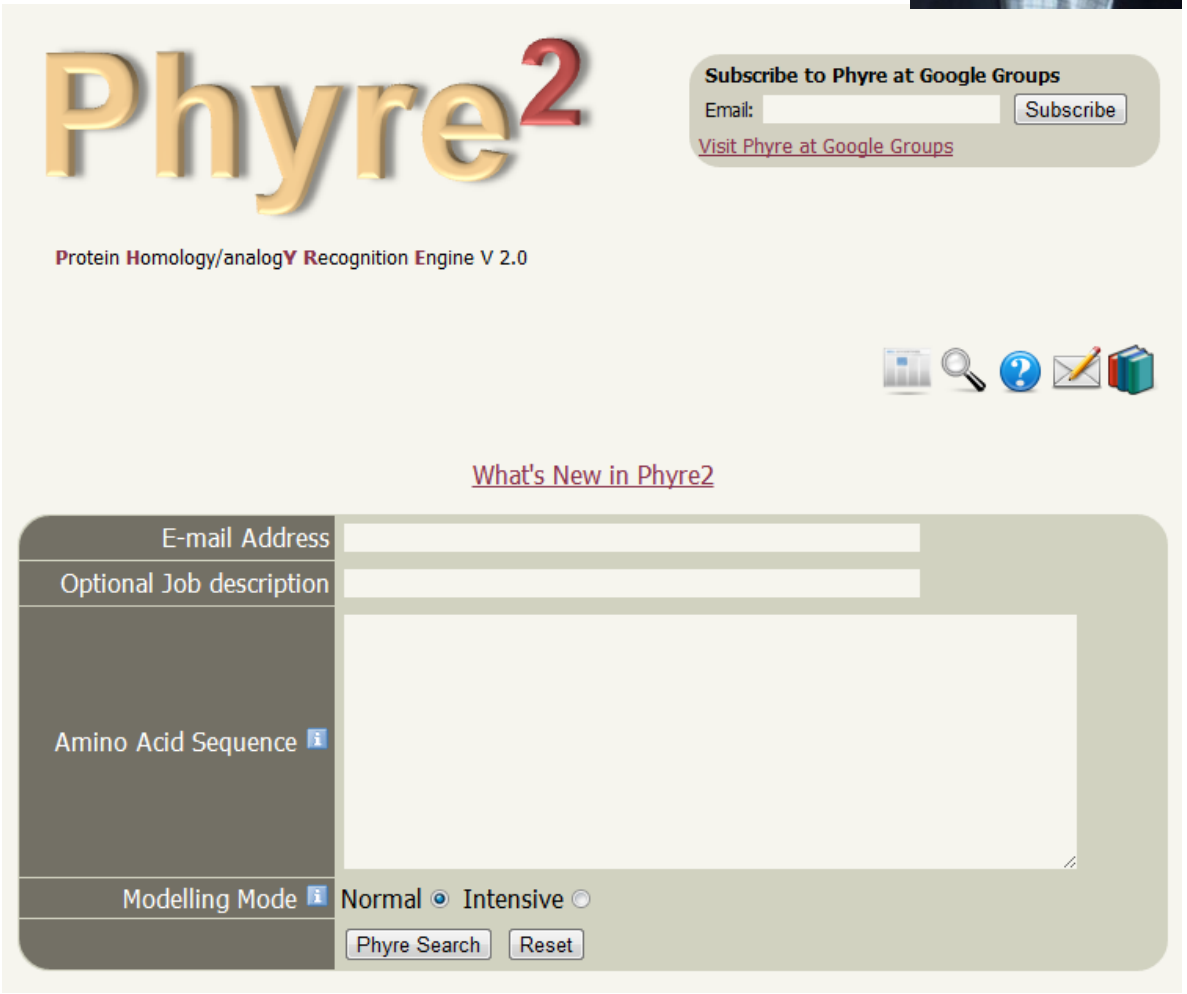
Typically 30 minutes, up to two hours

GenTHREADER Jones (1999) *J Mol Biol*

Homology modelling
Phyre

Mike Sternberg http://www.sbg.bio.ic.ac.uk/phyre2/

Kelley et al (2000) *J Mol Biol*
Kelley et al (2015)
Nature Protocols



Processing time can be hours

Homology modelling
Static solutions

Datasets of precomputed models / computations

Not flexible

Variable coverage

But you don’t have to wait

Homology modelling
MODbase

Andrej Sali

http://modbase.compbio.ucsf.edu/

ModBase: Database of Comparative Protein Structure Models

· Sali Lab Home · ModWeb · ModLoop · ModBase · ModEval · PCSS · FoXS · IMP · ModPipe ·

ModBase Home · ModBase Datasets for User:Anonymous · User Login · Help · News · Contact · Current Datasets

- General Information
- Statistics and Genome Datasets
- News
- Project Pages
- Authors and Acknowledgements
- Publications
- Related Resources

Please address inquiries to: modbase@salilab.org

MODBASE contains theoretically calculated models, not experimentally determined structures. The models may contain significant errors.

ModBase Search

ModBase is a database of comparative protein structure models, calculated by our modeling pipeline ModPipe.

Search

Search type: Model(Default) Display type: Model Detail (graphical)

To include the academic (comprehensive) dataset, go to 'Current Datasets'!

All available datasets are selected

Search by properties

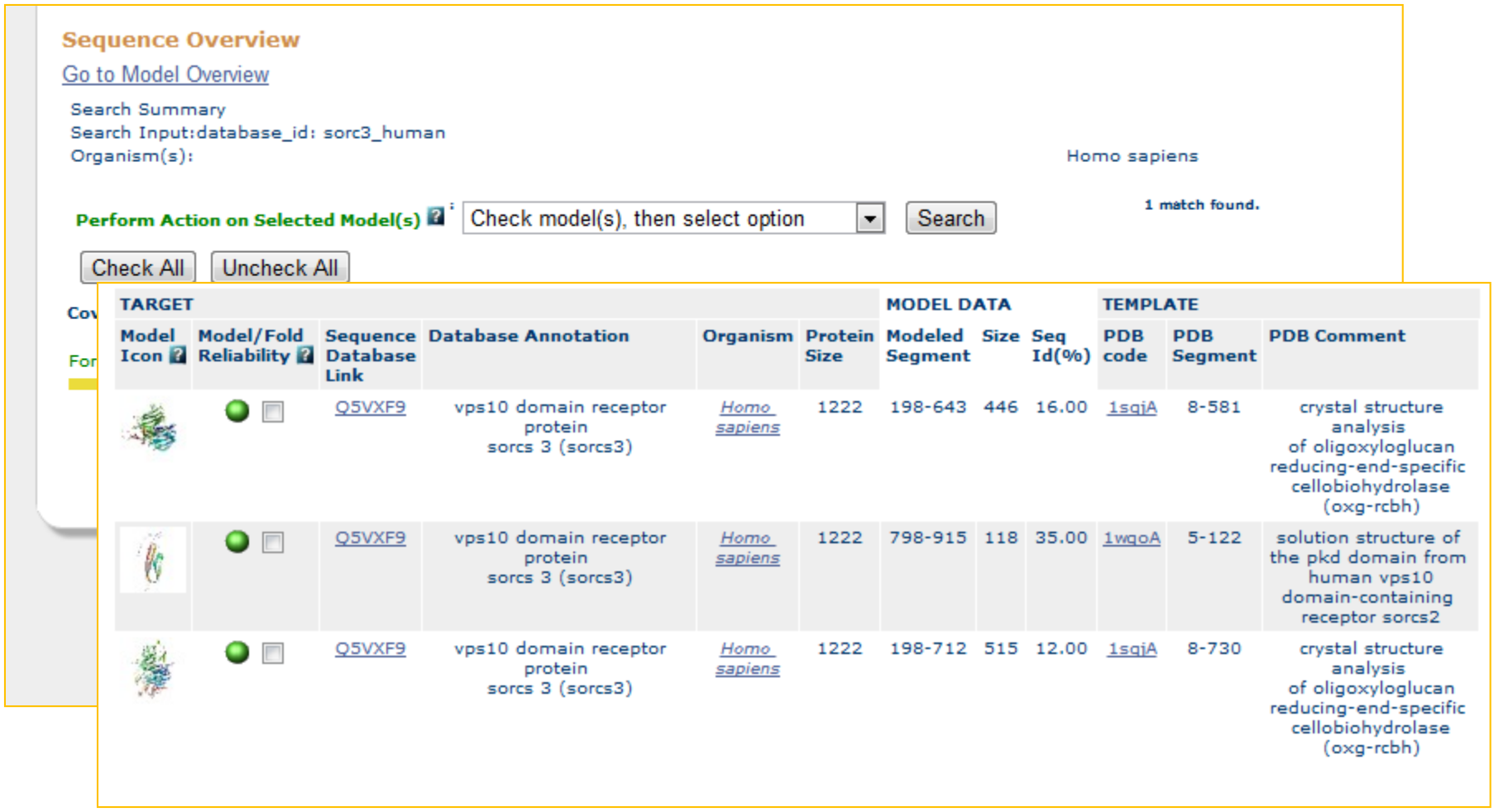
Property: Database Accession Number

Organism: ALL or

Advanced search

Pieper et al (2014) *Nucleic Acids Research*

Homology modelling
MODbase



Sequence Overview

Go to Model Overview

Search Summary
Search Input:database_id: sorc3_human
Organism(s): Homo sapiens

Perform Action on Selected Model(s) : Check model(s), then select option Search

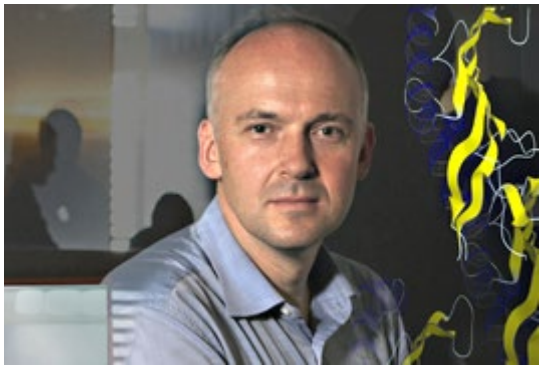
1 match found.

Check All Uncheck All

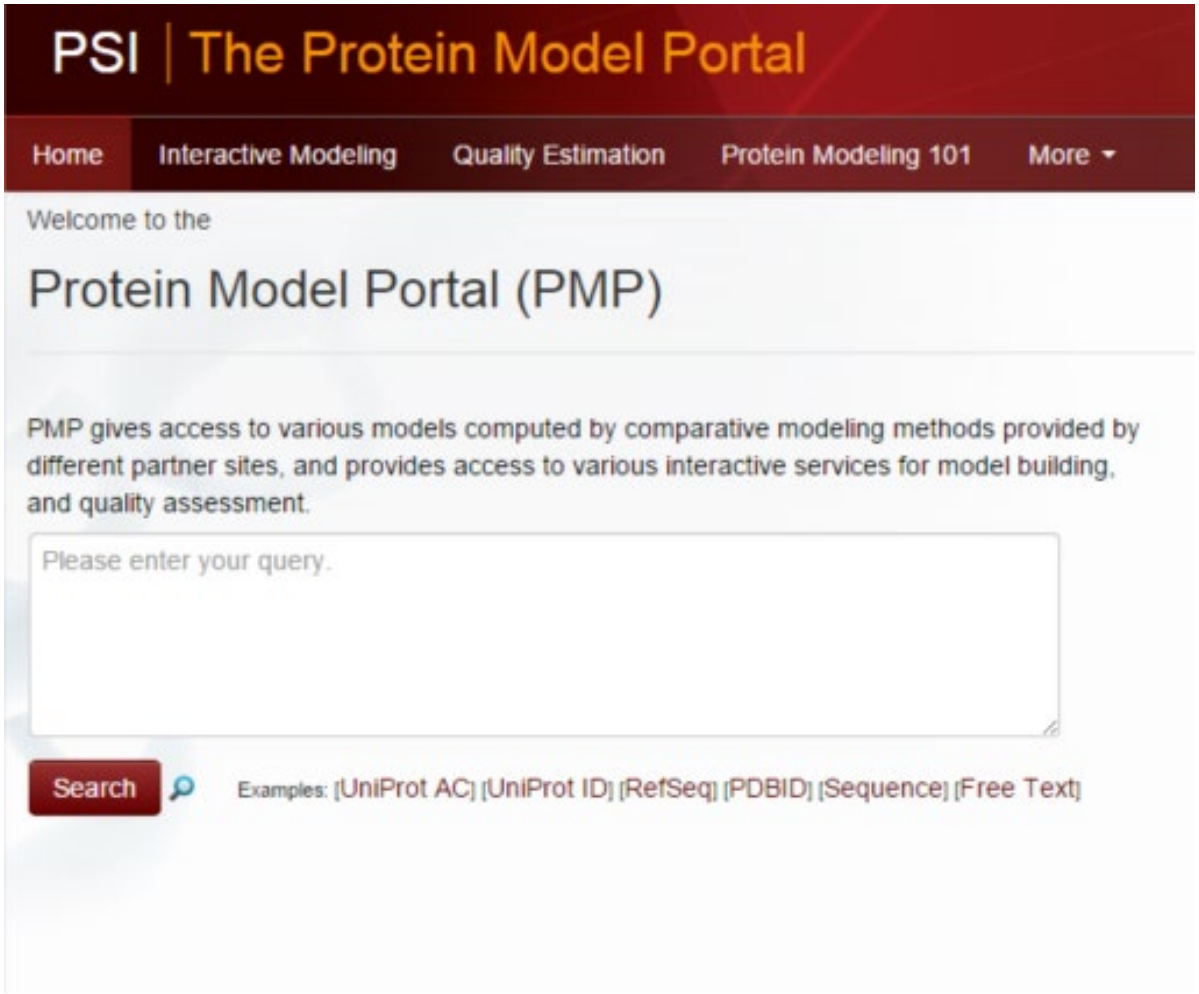
TARGET						MODEL DATA			TEMPLATE		
Model Icon	Model/Fold Reliability	Sequence Database Link	Database Annotation	Organism	Protein Size	Modeled Segment	Size	Seq Id(%)	PDB code	PDB Segment	PDB Comment
		Q5VXF9	vps10 domain receptor protein sorcs 3 (sorcs3)	Homo sapiens	1222	198-643	446	16.00	1sqjA	8-581	crystal structure analysis of oligoxyloglucan reducing-end-specific cellobiohydrolase (oxg-rcbh)
		Q5VXF9	vps10 domain receptor protein sorcs 3 (sorcs3)	Homo sapiens	1222	798-915	118	35.00	1wqoA	5-122	solution structure of the pkd domain from human vps10 domain-containing receptor sorcs2
		Q5VXF9	vps10 domain receptor protein sorcs 3 (sorcs3)	Homo sapiens	1222	198-712	515	12.00	1sqjA	8-730	crystal structure analysis of oligoxyloglucan reducing-end-specific cellobiohydrolase (oxg-rcbh)

Homology modelling

Protein Model Portal



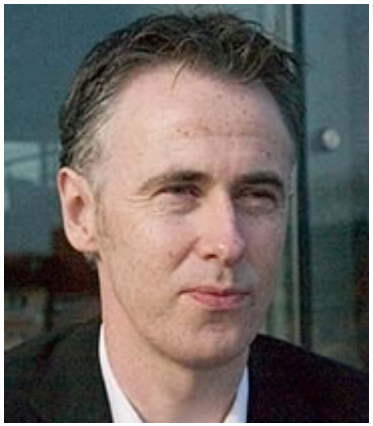
Torsten Schwede

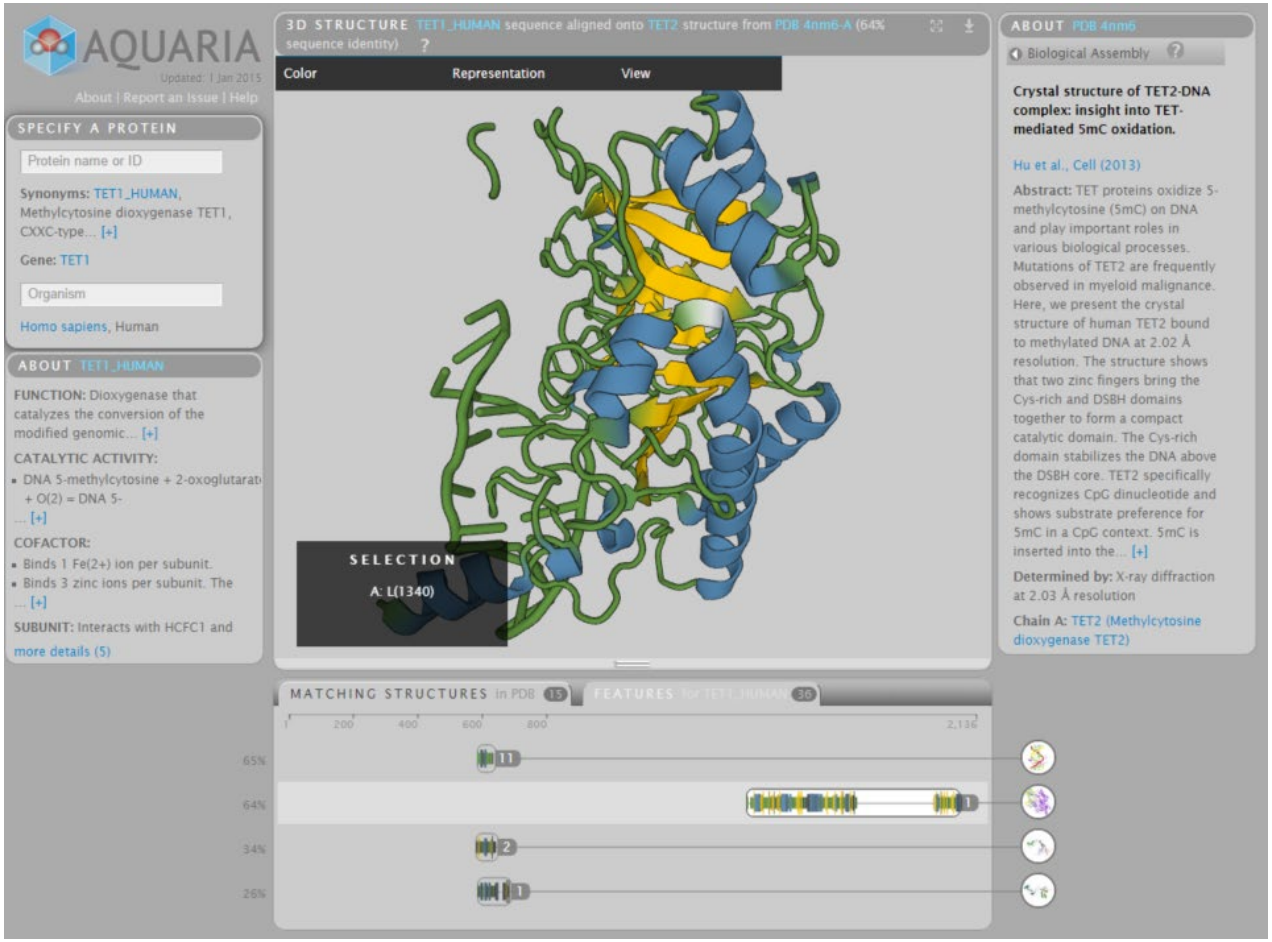


Haas et al. (2013) *Database*

Aquaria

Sean O'Donoghue http://aquaria.ws/





O'Donoghue et al (2015) *Nature Methods*

Domains

Protein domains are structural units (average 160 aa) that share:

Function
Folding
Evolution

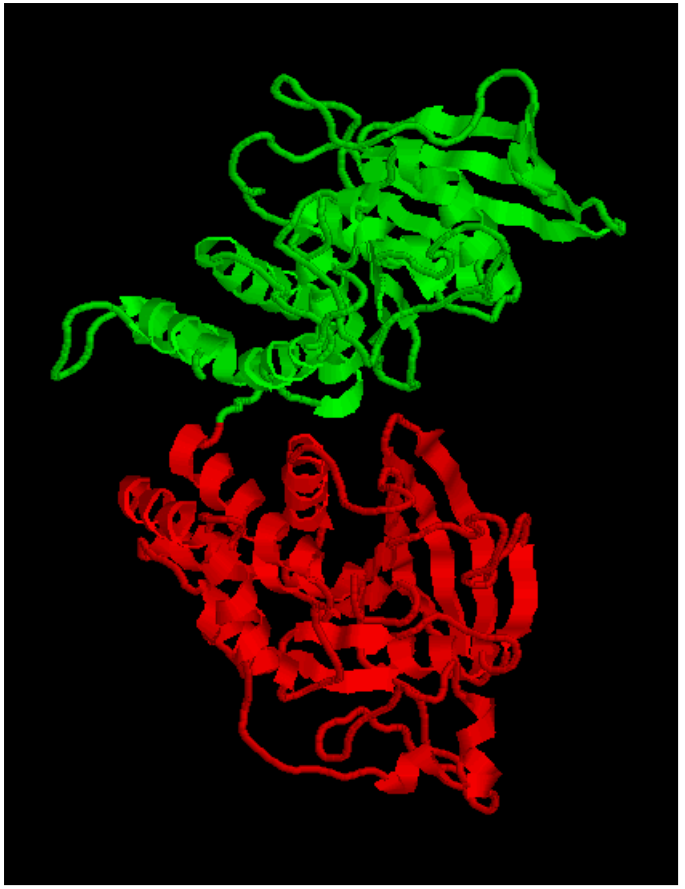
Proteins normally are multidomain (average 300 aa)

Domains

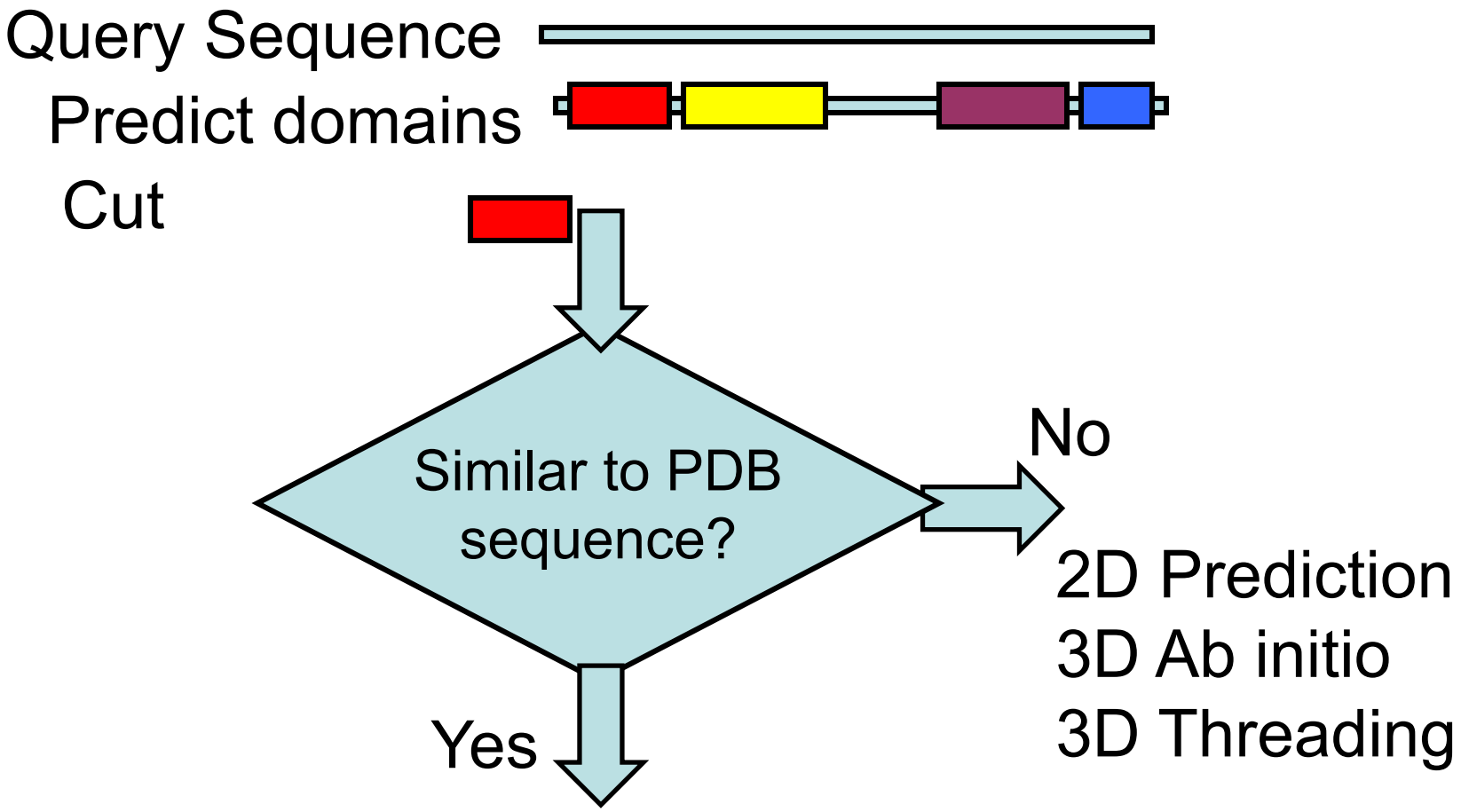
Protein domains are structural units (average 160 aa) that share:

Function
Folding
Evolution

Proteins normally are multidomain (average 300 aa)



Domains



3D Modeling by homology

3D structure prediction
Ab initio

Explore conformational space

Limit the number of atoms

Break the problem into fragments of sequence

Optimize hydrophobic residue burial and pairing of beta-strands

Limited success

3D structure prediction
Threading

I-Tasser: Yang Zhang &
Jeffrey Skolnick

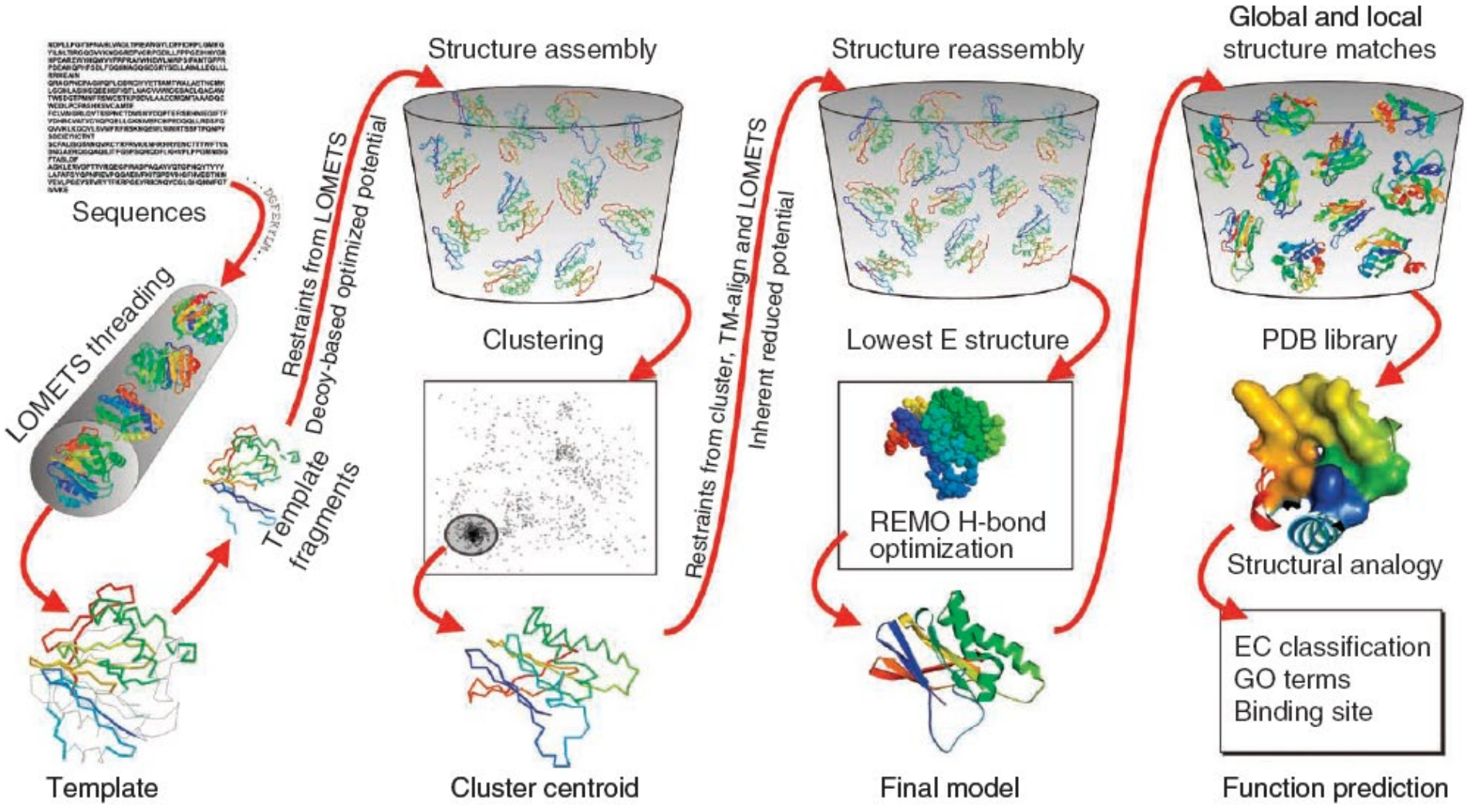
Fold 66% sequences <200 aa long of low homology to PDB

Just submit your sequence and wait… (some days)

Output are predicted structures (PDB format)

Lee and Skolnick (2008) *Biophysical Journal*
Roy et al (2010) *Nature Methods*
Yang et al (2015) *Nature Methods*

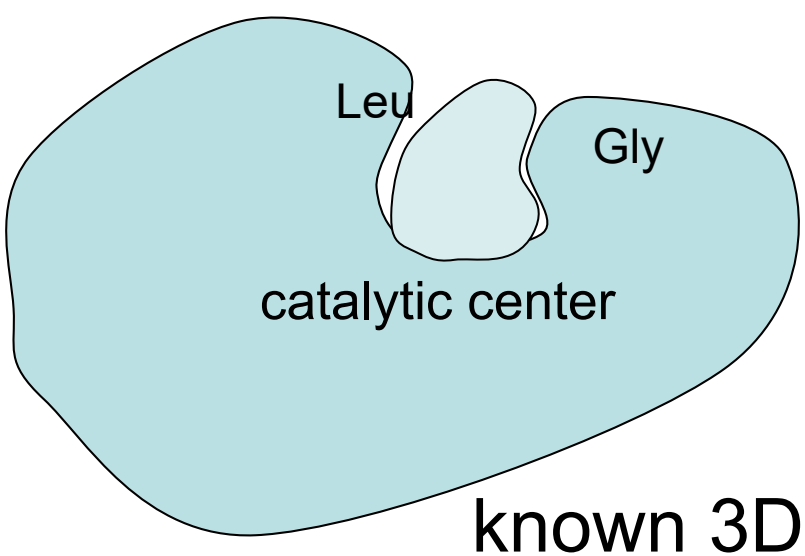
3D structure prediction
I-Tasser



Roy et al (2010) *Nature Methods*

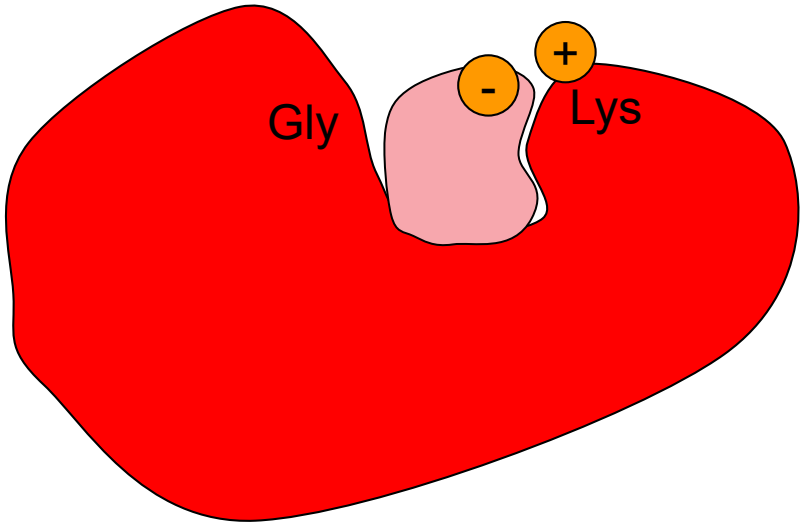
Structure-based drug-design

Find a molecule that fits a protein



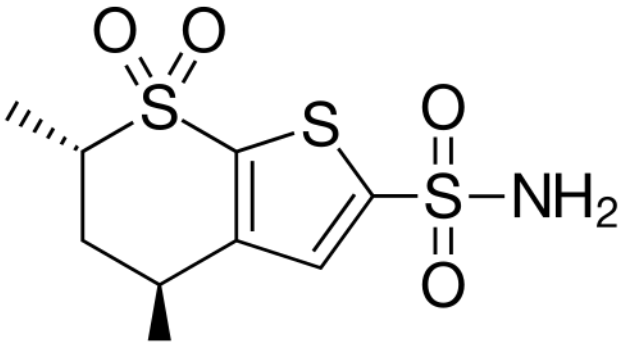
known 3D

or…



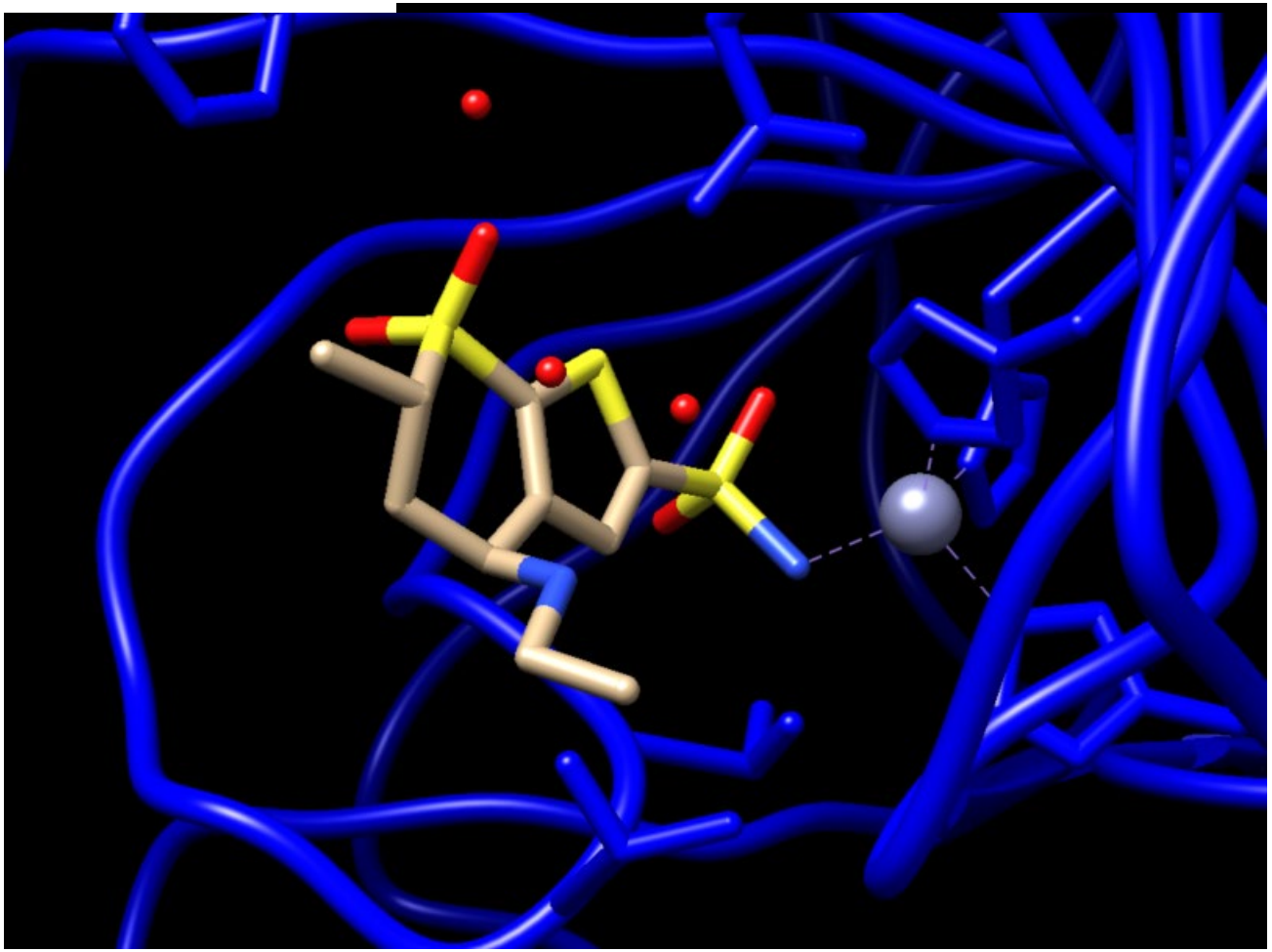
model 3D by homology

Dorzolamide



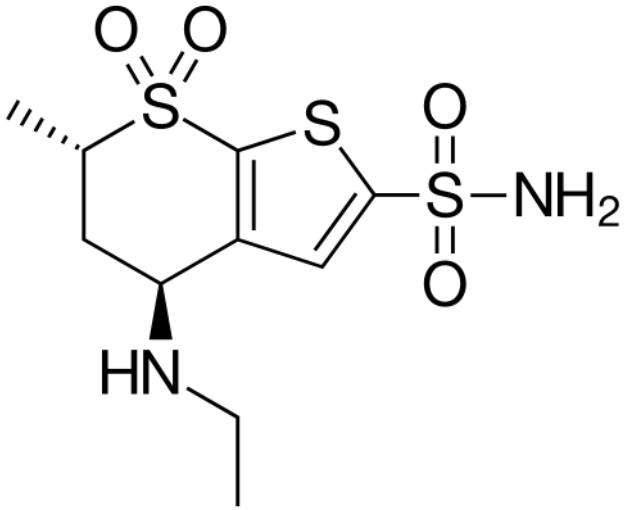
1st structure-based drug design: 1995

Carbonic anhydrase inhibitor



PDB: 1CIL

Dorzolamide



Decreases production of aqueous humour

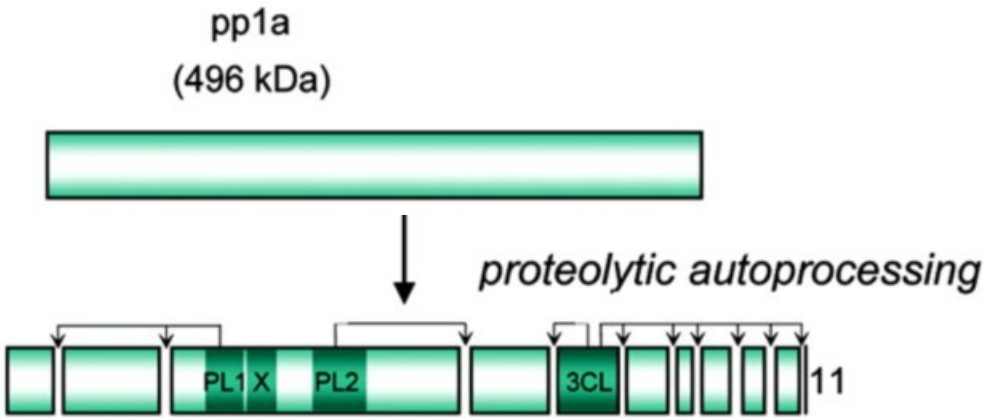
Reduces intraocular pressure

Eye drops to treat glaucoma and ocular hypertension

SARS example

SARS polyprotein as a drug target

Processed by proteinase 3CLpro



SARS example

Replicase polyprotein 1a Human SARS coronavirus: R1A_CVHSA

Polyprotein:
Analyze the domains in PFAM.

Can you see the peptidase domain?
What are its coordinates (amino acid positions)?

SARS example

Proteinase 3CLpro: activity

Cleaves at: Leu-Gln- [Ser, Ala, Gly]

P1 - P2

Can we design a drug to block this proteinase?

SARS example

We need the 3D structure of human SARS proteinase 3CLpro in complex with a target peptide.

Problem: X-ray doesn't work, cannot get crystals

We try with the proteinase 3CLpro from Porcine TGEV coronavirus (easier to culture).
It works! PDB: 1P9U

Can we do homology modeling of the human virus protein with the porcine virus protein?

SARS example

Proteinase 3CLpro: 3D structure from Porcine TGEV coronavirus in complex with a peptide.

Open in Chimera PDB: 1P9U.

How many chains do we have? What does it mean?

What is the sequence of the peptide?

How is the peptide bound to the protease?

SARS example

Proteinase 3CLpro: 3D structure from Porcine TGEV coronavirus in complex with a peptide.

Open in Chimera PDB: 1P9U.

Select chain F. Color it red.
Select chain H. Color it yellow.
Select zone (default 5 A)
Action > Atom/bonds > show

Cysteine 144 is important!

SARS example

Proteinase 3CLpro: 3D structure from Porcine TGEV coronavirus in complex with a peptide.

Open in Chimera PDB: 1P9U.

Three domains:
domain #1: 8-99
domain #2: 100-183
domain #3: 200-300

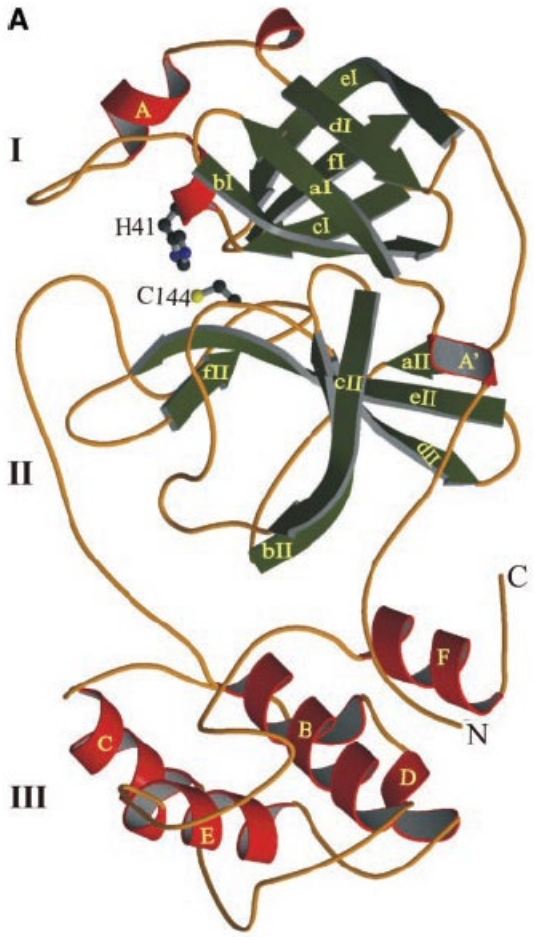
Binding site between domains #1 and #2

SARS example

Proteinase 3CLpro: 3D structure from Porcine TGEV coronavirus in complex with a peptide.

Open in Chimera PDB: 1P9U.

Catalytic diad: His 41, Cys 144



SARS example

We need the 3CLPro from SARS.

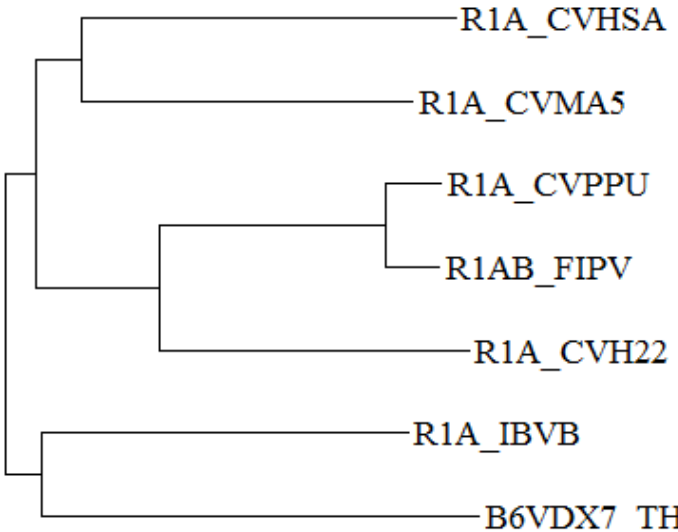
Are they similar enough so that we could do a model?

We did a multiple sequence alignment of several 3CLPro proteases TGEV, 229E, SARS and other related viruses:

From https://cbdm.uni-mainz.de/un17/
Open alignment PF05409_7seqs.txt using Bioedit or Jalview or Clustalw

SARS example

Phylogenetic tree:



Human SARS coronavirus (SARS-CoV)

Murine coronavirus (strain A59) (MHV-A59) (Murine hepatitis virus)

Porcine transmissible gastroenteritis coronavirus (strain Purdue) (TGEV)

Feline coronavirus (strain FIPV WSU-79/1146) (FCoV)

Human coronavirus 229E (HCoV-229E)

Avian infectious bronchitis virus (strain Beaudette) (IBV)

Thrush coronavirus HKU12 (isolate Grey-backed thrush/Hong Kong/HKU12-600/2007) (ThCoV-HKU12) (Thrush coronavirus HKU12-600)

Is Cysteine 144 conserved? (Note that in this alignment it is at position 119).
Anything else conserved around?

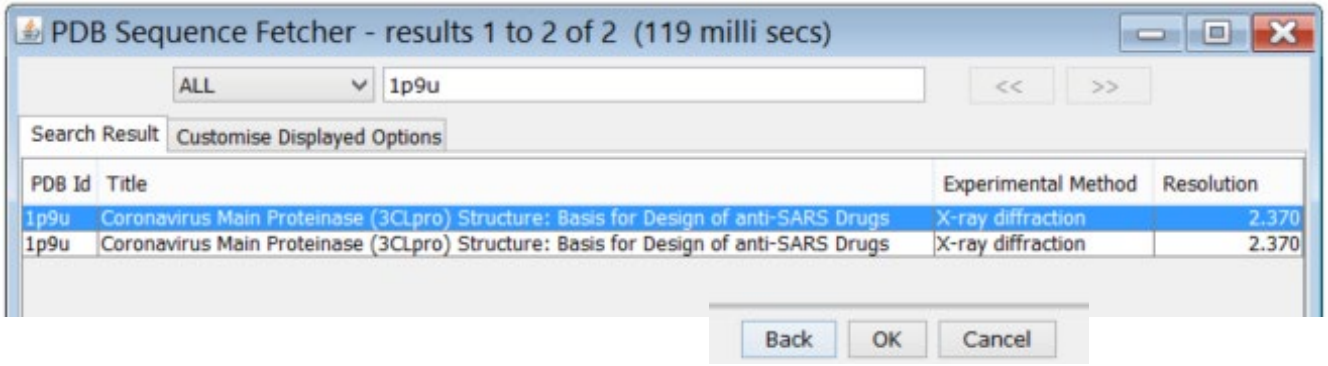
SARS example

Compare structures of model (SARS) and TGEV protein.

Open jalview (close all the windows that appear then:
Main Menu > Tools > Preferences > Visual:
Tick out option “Open file”)

Main menu > File > Fetch sequences > click Select database > click PDB

Type 1P9U, select one result and click OK

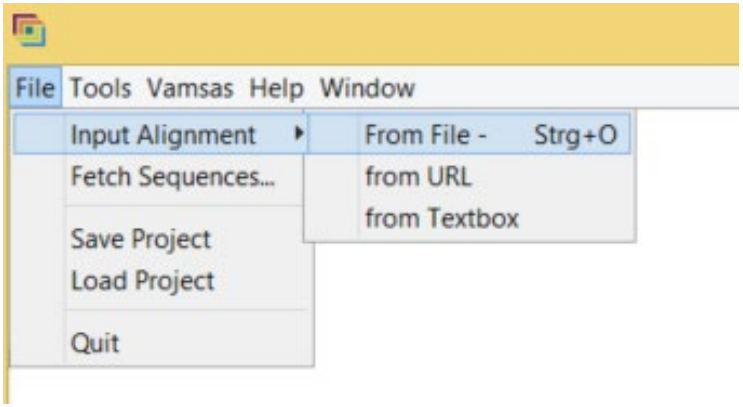


SARS example

Now we get the file with the model. It is not in PDB.
Find it at our course web page:
https://cbdm.uni-mainz.de/un17/

The name of the file is **1p9t.txt**
Right click on the link and save the file in your disk.

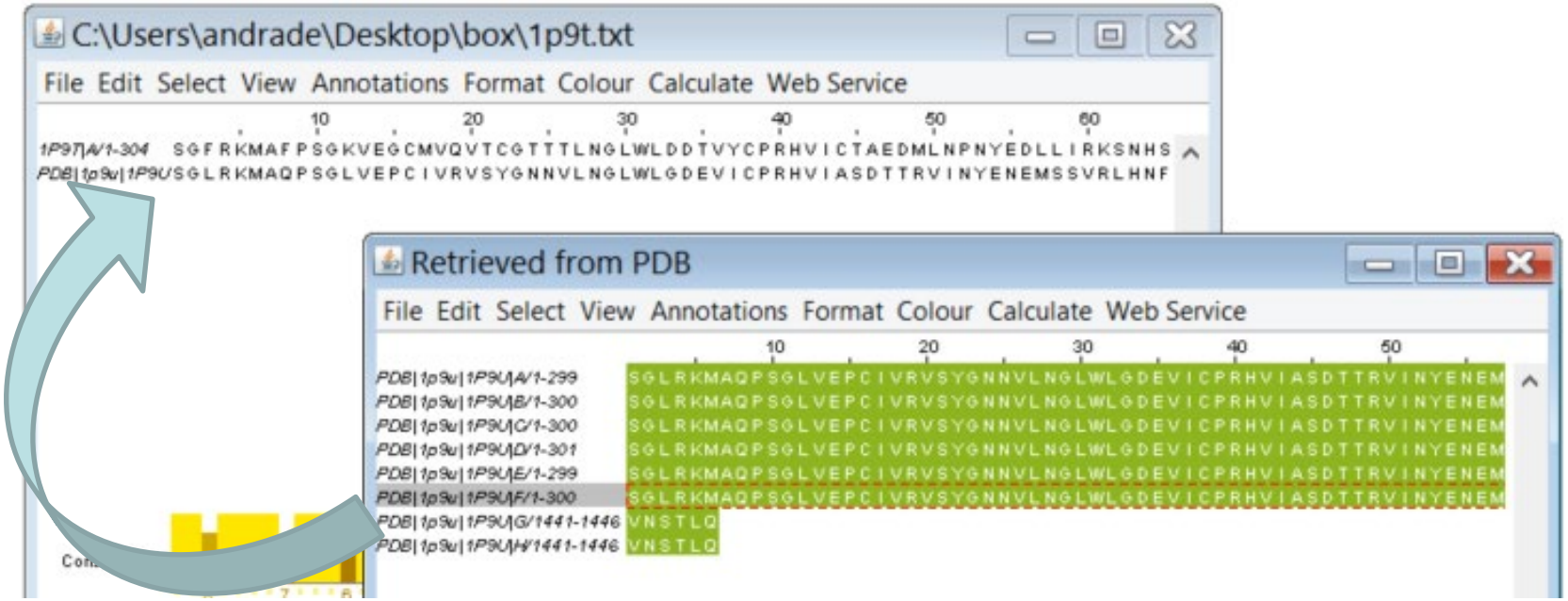
Then in jalview:
Main menu > File > Input alignment > From File



SARS example

Next let's put together both sequences.

Copy the 1P9U|F/1-300 to the window of 1P9T



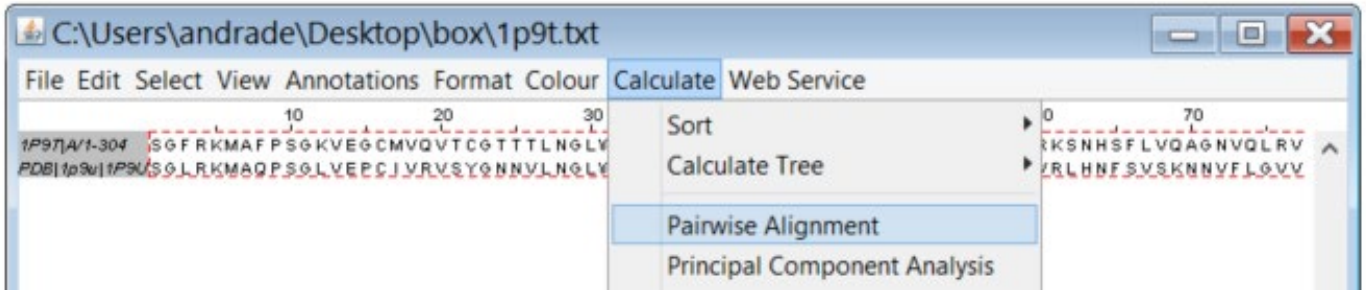
Look at the two sequences.

Do they look similar? Are they well aligned?

SARS example

To align the sequences: select the two sequences.

Menu (of this window!) > Calculate > Pairwise alignment



What's the percentage of identity between these two sequences?

Click the button "View in alignment editor"

Can you find gaps in the alignment: How many?

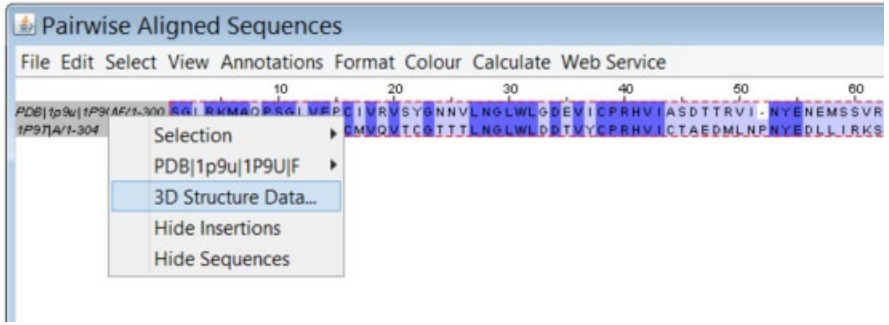
Color the alignment by "percentage identity"

SARS example

Now we will compare the two structures.

Select both sequences by clicking in the identifiers.

Right click on them.
Select 3D structure data…



Select both lines and click view.

What happened? :-)

SARS example

The story continues…

Berry 2015 - virtual screening